Conservation and Evolution of Microsatellites in Vertebrate Genomes

---

A thesis submitted in partial fulfilment

of the requirements for the Degree of

Doctor of Philosophy in Biological Sciences

in the University of Canterbury

by Emmanuel Buschiazzo

University of Canterbury

2008

There is enough light
For those who desire only to see,
But there is enough darkness for those
Of a contrary disposition

—BLAISE PASCAL
Pensées (1670)

…I claim
No private lien on the truth, only
A liberty to seek it, prove it in debate,
And to be wrong a thousand times to reach
A single rightness…

—MORRIS WEST
The Heretic (1969)

Biological intuitiveness and
Biological investigator empowerment
Need to take precedence over the current supposition that
Biologists should re-tool and become programmers when analyzing
Genome scale data sets

—SUDIR KUMAR & JOEL DUDLEY
Bioinformatics, 23-14 (2007)

Calvin: I think we've got enough information now, don't you?
Hobbes: All we have is one "fact" you made up.
Calvin: That's plenty. By the time we add an introduction, a few illustrations, and a conclusion, it will look like a graduate thesis. Besides, I've got a secret weapon that will guarantee a good grade! No teacher can resist this! A clear plastic binder!
Pretty professional looking, eh?
Hobbes: I don't want co-author credit on this, OK?

—BILL WATERSON
Calvin and Hobbes

AKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Neil Gemmell for giving me this great and unexpected opportunity to come to New Zealand and challenge my scientific skills with a PhD project on the evolution of microsatellites. He made things really easy right from the airport and understood very quickly that he could make me happy with just finding a few éclairs to munch on. I would also like to thank Assoc. Pr. Neil Gemmell for letting me the time to get confident with the subject, try out my ideas, while sharing his impressions on things and always giving me the right directions, which I sometimes realized a few weeks later only. Finally, I am greatful to Pr. Neil Gemmell for always having a door opened and time to read over my manuscripts despite his many occupations and responsabilities.

Other academics at the University of Canterbury have offered their help, advices and general support in exchange of nothing much, including, in no particular order: Tammy Steeves, for helping me go through some existential issues; Raaz Sainudiin, for giving me a fresh new way to look at my data and challenge me with unheard-of methodologies; Marie Hale, for giving me a VIP access to her PCR cycler; Bruce Roberston, Jack Heinemann, Sharyn Goldstien and Jason Tylianakis for their general help when needed.

Members of the Molecular Ecology Lab have obviously coloured my time at uni in many ways. My very special thanks go to Jonci 'Oldie Goldie' Wolff for aaaaall the good times and for his incredible support in the last hours of my PhD, and Antoine 'La Bête' Fouquet, for his much-appreciated frenchness and lightness mixed with deep discussions, commenting the news and many abberations of this world. These two guys have also rolled a lot of cigarettes for me! It was also an interesting time to spend with the rest of the microsatellite crew, which includes Angelika Merkel, Iris Vargas-Jentzsch and Andrew Bagshaw; we were a bunch of very different people, but we made it! The Gemmell's group

is reputably big, so I can't include everybody here, but for whatever good time we had, this was a great pleasure to have it with you all.

I have an immense gratitude towards all involved in the collection of mammalian DNA/tissue samples. Unexpectedly, this was one of the hardest tasks of my PhD, and I realized there how much Science needs generosity and friendship to advance.

Then there's Josephine Sarah Beck. Without the incredible and long-lasting effort that she managed to make for a project that was not hers, I would still be pipetting instead of writing these words; you know how much this means to me! She is an immense resource of happiness, giggles and enchantment! I thank her for sharing her time with me, teaching me the ins and outs of a responsible eater, giving me a window on what the kiwi life really is, and for soooo much more!

Many other people were associated with my great times in New Zealand, and I can only dream trying to achieve an exhaustive list of them. My very special thoughts would go to Camilo Rodriguez-Beltran, but if anybody who thinks should also have a place in here, just close your eyes and imagine all those nice letters forming your name materialize right here. I can't help being especially grateful to all those who put up with me turning into an asocial geek for a while. For those who didn't, I think I can understand.

Finally, there's this group of people that roams at the other side of the world, starting with my mother, Armelle Goubelle. If I achieved anything during my time at university, that's only thanks to her, her formidable support and willingness to give whatever she has to make me go a step further. My brothers Nicolas and Stephane Buschiazzo, and their family, have also participated in their way to my advancements, and I thank them for their many advices (what would you expect from big brothers?). I can't forget those people that gave their trust at crucial times and gave me the taste for research, especially Dr. Thomas Guillemaud and Dr. Michel Warnau. This is it, thank you!

TABLE OF CONTENTS

ABSTRACT

Microsatellites are strings of short DNA motifs (≤6 bp) repeated in tandem across genomes of both prokaryotes and eukaryotes. In 20 years, they became popular genetic markers, successfully employed in the field of genetic mapping and gene hunting, as well as to address various biological questions at the individual, family, population and species level. However, evolutionary and demographic inferences from microsatellite polymorphism are hampered by controversy and ambiguity in the mutational processes of microsatellite sequences. Drawing on new data from genome projects, I review in Chapter 1 the concept of a microsatellite life cycle, which hypothesizes that microsatellites follow a life cycle from birth, through expansion, contraction, death and potentially resurrection. To document and understand this integrative concept of evolution, which could help improve current models of microsatellite evolution, there is an implicit need to study the evolution of microsatellites above the species level. A prerequisite of such comparative studies is therefore to find microsatellite loci that are conserved between different species.

The near or full completion of many vertebrate genomes and their alignment against one another offer the ultimate approach to find genomic elements conserved over a large evolutionary scale. In Chapter 2, I present a new comprehensive method to find conserved microsatellites in whole genomes. Using the multiple-alignment of the human genome against those of 11 mammalian and five non-mammalian vertebrates, I examine the genomewide conservation of microsatellites, and challenge the general assumption that microsatellites are too labile to be maintained in distant species. In Chapter 3, I present similar results using the alignment of the newly sequenced platypus genome against those of three mammals, the chicken and the lizard, and incorporate these data into the framework created by the 17-genome analysis. This enlarged dataset was ground for

attempting to reconstruct a vertebrate phylogeny from the presence/absence of microsatellites in the different genomes. Maximum parsimony analyses resulted in a tree much similar to that of the current view of the vertebrate phylogeny, while Bayesian analyses showed some discrepancies. This work opens a way for novel theoretical developments regarding the inference of ancestral states of microsatellites. In Chapter 4, I show how knowledge on conserved microsatellite sites can help for the development of a set of comparative primers useful across the Mammalia; implementing a similar protocol, nine conserved dinucleotide repeats were genotyped in 20 unrelated individuals of 18 species (nine sister species) encompassing the mammalian phylogeny, including marsupials and monotremes, and four microsatellites were sequenced in 4 individuals per species. My results emphasize conserved microsatellites as a new resource for genetic mapping and population studies. Finally, in Chapter 5, I recount the unexpected extent of structural change among mammalian orthologous microsatellites, including change of complexity, motif replacement and overall length variability. Altogether, these findings provide a comprehensive framework that may help in many areas of research, including molecular ecology, genome mapping, population genetics, and genome and microsatellite evoluion.

LIST OF FIGURES

LIST OF TABLES

Chapter 1

1 Introduction*

* Buschiazzo E and Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28: 1040-1050

1.1 Microsatellites

1.1.1 Definition

Vertebrate genomes can be broadly regarded as patchworks of unique and repeat sequences. Two distinct types of repeats are found, depending on whether the repeat units are dispersed (interspersed repeats) or clustered together (tandem repeats). Satellite DNA was the first of the tandemly repeating sequences to be discovered, and was so named due to its appearance as satellite bands in ultracentrifuge density gradients of complex eukaryotic genomes (Corneo et al. 1967). By extension, the term 'satellite' has been declined when smaller classes of tandem repeats were identified. Intermediate-sized repeats were called minisatellites (Jeffreys et al. 1985) and, consequently, the smallest sized repeats were dubbed microsatellites (Litt and Luty 1989; Tautz 1989).

Microsatellites, defined as strings of short motifs (1-6 bp) tandemly repeated, are found in the genomes of prokaryotic and eukaryotic organisms (Hancock 1999). The Human Genome Sequencing Consortium (Lander et al. 2001) estimated that microsatellite sequences comprise ~3% of the human genome (Figure 1.1), but significant variation in microsatellite content is observed between species (Tóth et al. 2000; Dieringer and Schlötterer 2003; Warren et al. 2008).

Microsatellites are highly polymorphic sequences; they mutate through addition or removal of repeat units from the array, but point mutations can also occur within the array and create an imperfection. So-called perfect and imperfect microsatellites are thus generally distinguished.

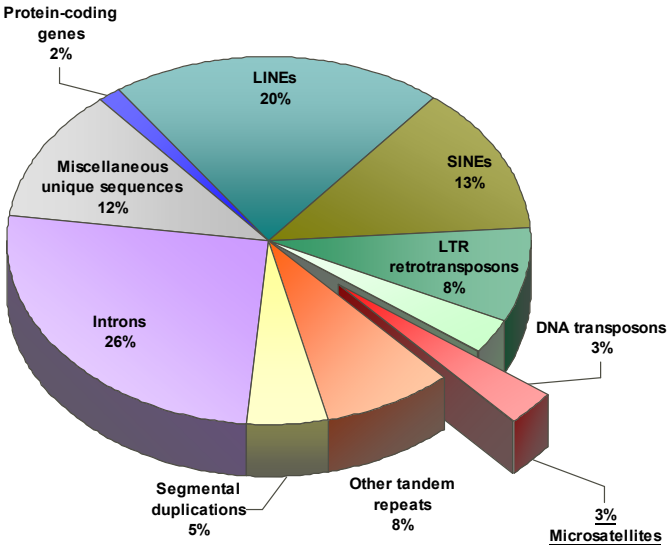


Figure 1.1: Composition of the human genome.

1.1.2 Distribution and motif preference in genomes

With the availability of genomic data, it became clear that the distribution of microsatellites in genomes was ubiquitous but non-random (Metzgar et al. 2000; Tóth et al. 2000; Katti et al. 2001; Li et al. 2002; Subramanian et al. 2003; Zhang et al. 2004; Cruz et al. 2005; La Rota et al. 2005; Lawson and Zhang 2006; Grover et al. 2007; Kim et al. 2008). For example, fewer microsatellites, mostly tri- and hexanucleotide repeats, were found in coding regions of eukaryotic genomes (Metzgar et al. 2000).

In addition, while motif abundance is similar between chromosomes of a same species, it was found to vary significantly between species (Subirana and Messeguer 2008). In the human genome, $(AC)_n$ repeats make up for 50% of dinucleotide repeats, whereas $(CG)_n$ repeats contribute for only 0.1% (Lander et al. 2001). By contrast, $(AG)_n$ and $(AT)_n$ repeats add up to 98% and 88% of dinucleotide repeats in the *Arabidopsis thaliana* and rice genomes, respectively (Lawson and Zhang 2006).

These various non-random, species-specific genomic patterns have led to many questions, often still partially unresolved, about how microsatellites arise, how they are maintained and ultimately how functionally important they might be.

1.1.3 Functions

Certainly, some microsatellites have functions in and/or influence on genomes, e.g. in regulating gene expression (reviewed in Li et al. 2002 and Kashi and King 2006). For example, a dinucleotide repeat seems to affect mating behaviour of voles (Hammock and Young 2005), although this claim has recently been questioned (Fink et al. 2007). It is also debated whether microsatellites are involved in recombination; although the general consensus is that there is no effect (Kayser et al. 2000), it has been shown for example that microsatellites (1-3 bp) were twice as frequent in recombination hot spots than cold spots of the yeast genome (Bagshaw et al. 2008). These contradictions stress the need for more research to unravel the function of microsatellites in genomes.

In addition, an ever increasing number of unstable repeats, mostly coding trinucleotide repeats, are implicated in ~30 human hereditary disorders (Mirkin 2007), while others are associated with colorectal, endometrial, and various other cancers (Woerner et al. 2006).

1.1.4 Applications

Despite an uncertainty around the evolutionary dynamics of microsatellites, their outstanding abundance and high variability have resulted in microsatellites emerging as the

genetic marker of choice over the last decade (Chambers and MacAvoy 2000; Schlötterer 2004). Extensively used in genome mapping and gene hunting, microsatellites have also helped to address an impressive range of biological questions, from the level of the individual (identity, sex), the family (parentage, relatedness), the population (genetic structure, epidemiology) and species (phylogenetics).

1.2 Microsatellite evolution

All genetic markers used to assess genetic distance (e.g. in population genetics, phylogeography and phylogenetics) depend on the knowledge of the mutation processes that generate their variation, and on the robustness of the underlying estimates of mutation model parameters, such as the mutation rate or directionality.

A wide range of models have been proposed to explain the mutational dynamics of microsatellites (Box 1.1). It is arguable whether there is one possible best model to explain variation at microsatellite loci. A fair question to ask is whether the choice of model really matters, as biologists might feel that the resolving power of microsatellites outweighs the alleged simplicity of the Stepwise Mutation Model (SMM) and the Two-Phase Model (TPM). But for most, the oversimplicity of assumptions contained in present theories cannot be ignored when estimating genetic distance, especially when high divergence is envisaged. Although it is unclear how much complexity can be ignored while still closely reflecting empirical observations, the incorporation of most of the known features of microsatellite dynamics is required to aim at the challenging development of an integrative and realistic body of theory.

Box 1.1: Models of microsatellite evolution

Infinite Allele Model

The simple infinite allele model (IAM) assumes that each mutation creates a new allele in the population (Kimura and Crow 1964). However, the forward-backward mutation process at microsatellite loci ultimately results in the creation of alleles identical in state, a condition referred to as size homoplasy. Only the unusual dynamics of compound/complex microsatellites seem to be described best by the IAM.

Stepwise Mutation Model

Under the stepwise mutation model (SMM), mutations accrue via the addition or deletion of a single repeat at a time (Ota and Kimura 1973). Gains and losses occur at equal frequency and at a rate independent of allele size. Various estimators of genetic distance based on the SMM have been developed for phylogenetic and demographic applications. Even though the SMM is adequate when closely related populations are considered, this simplistic model may be inadequate when a critical level of divergence is reached.

Two-Phase Model

The two-phase model (TPM) is an extension of the SMM that allows for infrequent multistep mutations: one-step mutations are more likely to occur and follow the SMM, whereas the magnitude of multistep mutations follows a truncated geometric distribution (Di Rienzo et al. 1994). Some contention has been raised around studies that found better fits with the TPM than with the SMM, as they used allele size scored from polymerase chain reaction (PCR) product length, and thus could not account for length change mutations in the flanking regions.

Biased Mutational Process Models

A number of sophisticated models have been proposed to explain the many complexities of microsatellite mutational dynamics, e.g. dependence of the mutation rate on allele length and on the number of point mutations (Kruglyak et al. 1998), mode and tempo of expansion and contraction events (Sainudiin et al. 2004), directional bias (Calabrese and Durrett 2003) and upper length constraint (Garza et al. 1995). However, these models have not been routinely applied to empirical data.

1.2.1 The life cycle concept of microsatellite evolution

A wide array of experimental approaches has been used to study microsatellite dynamics (Vargas-Jentzsch et al. 2008) in many different species, thus producing a large amount of data. Drawing on these data, it has been hypothesised that microsatellites follow a life cycle from birth, through expansion, contraction, death and potentially resurrection.

Although Messier and co-workers (1996) and Gordon (Gordon 1997) have already used the semantic of 'birth' to debate how a microsatellite locus appeared during primate evolution, Amos (Amos 1999) first proposed a life cycle pattern of evolution for microsatellite loci. Later, Taylor and co-workers (1999) observed the degeneration, or 'death', of a microsatellite locus, and Chambers and McAvoy (Chambers and MacAvoy 2000) made a step forward in completing the life cycle by suggesting that a dead microsatellite locus could potentially resuscitate (Figure 1.2A). Unfortunately, it seems that an overly vague conceptual framework, caused by a lack of thoroughness, a mathematical gap, the difficulty to compare results from different studies, and/or simply the lack of available supporting data, nipped the concept of a microsatellite life cycle in the bud. It could be argued that this situation has led to a deficit in recent progress on microsatellite evolution. In particular, the development of an integrative point of view on the evolution of microsatellites encapsulating new data from the genome projects has been missing.

While currently not widely recognised, the life cycle concept has the advantage of outlining the mutational processes and biases observed at microsatellite loci in a dynamic evolutionary framework, providing ground for the future development of a realistic model of microsatellite evolution.

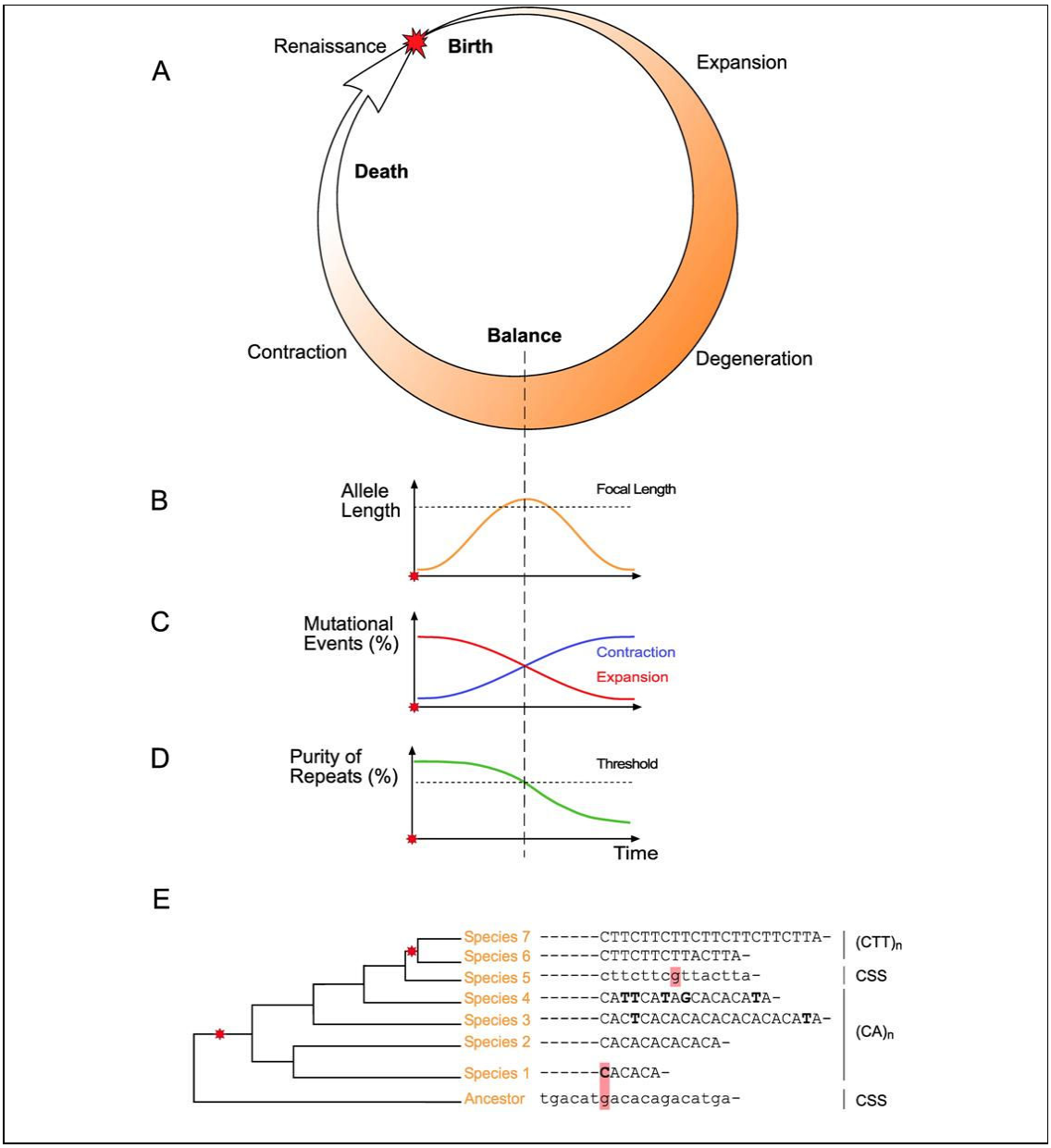


Figure 1.2: Hypothesised biology of a microsatellite locus. (A) Schematic life cycle and **(B–D)** variations over time in: **(B)** average allele length, **(C)** relative occurrence of repeat expansion and contraction events and **(D)** purity of repeat (proportion of perfect repeats in the array). **(E)** Data superimposed on a phylogeny allow direct observation of a locus life cycle (CSS:Cryptically Simple Sequence). After initiation, a microsatellite expands and, with increasing length, interruptions by point mutations occur and affect the expansion mutation rate (see text). The repeat array reaches an upper length limit where expansion and contraction mutations are in balance. While long deletions then occur predominantly and reduce the microsatellite in size, the continuing accumulation of imperfections breaks the array and decreases the rate of slippage. Both events lead to the fading of the microsatellite, but another locus may emerge from the remaining scramble of unique sequences. Time scale, upper allele length limit, dynamics of expansion versus contraction mutation events, and purity threshold are quantitatively unknown and are probably variable among loci.

1.2.2 Birth and maturation

The genesis of microsatellites remains a matter of debate (Wilder and Hollocher 2001). Published observations of such events generally result from opportunistic circumstances rather than from systematic approaches to investigate microsatellite genesis (e.g. Messier et al. 1996). Despite this, the genesis of microsatellites in genomes appears to be non-random, with an imbalance between the mechanisms that promote and those that prevent the initiation of microsatellites.

Current data suggest two alternative, but not mutually exclusive, hypotheses to explain microsatellite genesis. These hypotheses suggest that microsatellites arise either spontaneously from/within unique sequences (Messier et al. 1996) (*de novo* microsatellites), or that they are brought about in a primal form into a receptive genomic location by mobile elements (Wilder and Hollocher 2001) (adopted microsatellites).

De novo microsatellites are assumed to arise via the creation of a proto-microsatellite, i.e. a short intermediate stage with as few as 3 or 4 repeat units, within cryptically simple sequences, which are defined as a scramble of repetitive motifs lacking a clear tandem arrangement (Hancock 1999). Proto-microsatellites were first thought to originate from base substitution(s), e.g. GAC**G**CACG→GAC**A**CACG, and to be the substrate for further expansion (Messier et al. 1996 but see Gordon 1997). Recently, it has been argued that proto-microsatellites formed frequently without the preceding nucleotide substitution, but rather through indel events (Dieringer and Schlötterer 2003). This is supported by the observation that insertions tend to copy adjacent bases (e.g. GCAT→GCA**CA**T), creating a proto-microsatellite (Zhu et al. 2000; Nishizawa and Nishizawa 2002). Zhu et al. (2000)

showed that the proportion of substitution events relative to insertion events is length-dependent, with substitutions being the dominant source of new two-repeat loci, while all new 4-5 repeat mutations come from insertions. While this is a plausible model, the use of the Human Gene Mutation Database may limit the generality of this study, as these findings are based on sequences subject to specific selective forces and, in any case, not representative of the whole genome. This is crucial as it has been demonstrated in mammals that substitution rates vary within genomes (Ellegren et al. 2003), which in turn would affect the mutational dynamics of nearby microsatellites (Santibáñez-Koref et al. 2001). A survey aimed to be representative of the whole genome, across a range of taxa, would be worthwhile to understand which type of mutation is responsible for the rise of microsatellites from cryptic sequences. The recent mapping of INDELs in the human genome offers such an opportunity (Mills et al. 2006).

The alternative model is that microsatellite sequences are adopted from other genomic regions via a number of transposable elements (TEs) found in abundance in eukaryotes and thought to shape genome evolution (Kazazian 2004). These TEs may contain one or more sites predisposed to the formation of microsatellites, hence favouring the dispersal of microsatellites in genomes. So far, the focus of this model has been on non-autonomous and non-LTR retrotransposons (Kazazian 2004), respectively Short and Long Interspersed Elements (SINEs/LINEs), as potential source for proto-microsatellites (Wilder and Hollocher 2001). Retro-pseudogenes may also be a source for proto-microsatellites, although there is little evidence to date to support their role in microsatellite origin (Nadir et al. 1996).

In mammals, microsatellites have long been demonstrated to be associated with *Alu* (Arcot et al. 1995; Nadir et al. 1996; Batzer and Deininger 2002) and L1 elements (Duffy

et al. 1996), respectively mammalian commonest SINE and LINE. Strikingly, at least 54% of human Y microsatellites are thought to have originated from these retrotransposons, and this figure is likely to be higher in autosomes (Kayser et al. 2004). The poly-(A) tract at the 3'-end of mammalian SINEs/LINEs provides a site amenable to reverse transcription errors leading to the genesis of A-rich proto-microsatellites and their expansion, should slippage occur. This widespread process would explain at least partially why ($A_{2-5}N$) motifs regularly compose the most abundant microsatellites in eukaryotes (Tóth et al. 2000). The fact that avian SINE/LINE elements do not terminate in poly-(A) tails is a likely explanation, yet certainly not the sole one, for the generally low frequency of microsatellites in birds (Primmer et al. 1997). A similar explanation may also account for highly abundant plant LTR retrotransposons and the lack of association between microsatellites and repetitive DNA in plants (Morgante et al. 2002). However, microsatellites may also consistently rise in the same locations outside the 3' poly-(A) tail, like in the dipteran "microsatellite initiating mobile elements" (*mini-me,* Wilder and Hollocher 2001) or in *Alu* elements associated with Friedreich ataxia (Clark et al. 2004). In such cases, the substitution/indel-slippage model of microsatellite birth (Dieringer and Schlötterer 2003) would act after transposition has occurred, unless a proto-microsatellite is already present and ready for slippage and further expansion.

There are one or more mechanisms that result in the birth of a microsatellite at any particular place in the genome, but also one or more mechanisms that must fail if they are to be maintained or spread, assuming that a control on microsatellites is required. The differential genomic distributions of microsatellites illustrate both the variability in birth rate throughout the genome and also that selective forces are acting against microsatellite birth in specific regions (Tóth et al. 2000; Katti et al. 2001; Metzgar et al. 2002;

Subramanian et al. 2003; Zhang et al. 2004; Cruz et al. 2005), e.g. in coding regions where microsatellites with tri- and hexanucleotide repeat motifs only are found more frequently than expected by chance (Metzgar et al. 2000; Tóth et al. 2000). This is presumably because addition or deletion of such motifs does not disrupt the reading frame in coding regions (Metzgar et al. 2000). Alternatively, some cellular factors, such as RNAi in *Caenorhabditis elegans* (Robert et al. 2005), antiretroviral resistance proteins in human (Bogerd et al. 2006), or cytosine methylation (Yoder et al. 1997) may be acting against TE dispersal, hence preventing TE-associated microsatellites from spreading in genomes. The recent finding that transposon-free regions are maintained throughout mammalian (Simons et al. 2006) and even vertebrate evolution (Simons et al. 2007) offers a straightforward opportunity to test a direct association between the presence/absence of microsatellites and that of TEs.

Once tandem duplications are generated, these short simple sequences may be prone to slippage (Rose and Falush 1998), i.e. the proto-microsatellite has graduated to the mature phase of its life cycle. While some authors debate the existence of a threshold size for initial expansion (Pupko and Graur 1999; Perez et al. 2005), others have proposed threshold sizes ranging from 4 to 8 repeats (Rose and Falush 1998; Sibly et al. 2001; Lai and Sun 2003; Shinde et al. 2003). Four or more repeats seems to be a workable minimum, as expansion of microsatellites with two repeats only might be due to the indel-duplication model explained above, whilst anything of 4 or greater might more reasonably be subject to slippage, e.g. (Primmer and Ellegren 1998). When slippage does occur, tandemly duplicated repeats will be added and will expand the array in length (Rose and Falush 1998).

1.2.3 Growth

Mutation rate in microsatellites is on average high, ranging from 10^{-7} to 10^{-3} mutations per locus per generation in eukaryotes (Primmer et al. 1996; Schug et al. 1997; Kruglyak et al. 2000; Vigouroux et al. 2002; McConnell et al. 2007). However, these figures are only a static snapshot of a much more dynamic picture, with a complex heterogeneity of mutational events frequently observed at allele-, locus-, individual- and/or taxon-levels (Di Rienzo et al. 1998; Schlötterer et al. 1998; Colson and Goldstein 1999; Anderson et al. 2000; Ellegren 2000; Makova et al. 2000; Neff and Gross 2001; Webster et al. 2002; Shao et al. 2005; Lia et al. 2007; Lopez-Giraldez et al. 2007; Kelkar et al. 2008). For example, there has been considerable debate as to why human dinucleotide repeats are longer and more polymorphic than their orthologues in chimpanzees (reviewed in Webster et al. 2002 and Vowles and Amos 2006). The consensus view is that the course and rate of microsatellite mutation are highly affected by a number of more or less intercorrelated factors that can be classified into 5 groups: mutation mechanisms, nature of microsatellite, genomic context, individual biological context and selective influences (Figure 1.3).

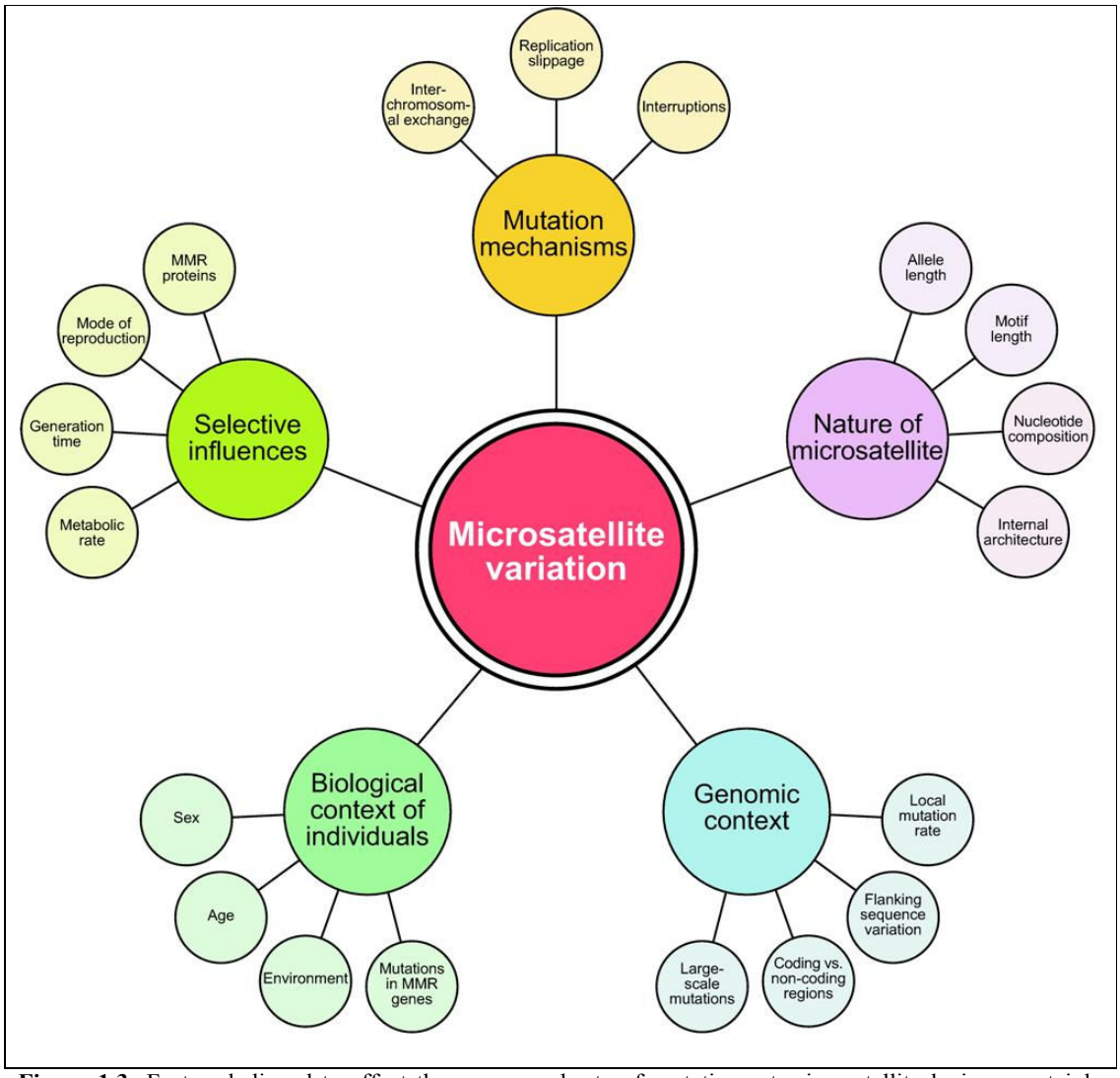


Figure 1.3: Factors believed to affect the course and rate of mutations at microsatellite loci are certainly intercorrelated and have a varying degree of influence.

Mutation mechanisms. Following their initiation, microsatellites are thought to vary in length (Figure 1.2B) by a stepwise mechanism of gain and loss via two mechanisms, namely replication slippage and interchromosomal exchange (reviewed in Ellegren 2004). The former involves dissociation of the replicating DNA strands followed by an out-of-register realignment that results in either a gain or a loss of generally one repeat unit, depending on whether a loop is formed on the nascent or the template strand, respectively (Levinson and Gutman 1987). Although the majority of these slippage events involve single-step mutations (Weber and Wong 1993; Amos et al. 1996; Brinkmann et al. 1998;

Kayser et al. 2000; Leopoldino and Pena 2002; Gusmão et al. 2005; Nikitina et al. 2005), i.e. addition/deletion of a single repeat unit, high incidences of multi-step events, involving addition/deletion of multiple repeat units, have occasionally been recorded: 46.2% in lizards (Gardner et al. 2000) and 50% in freshwater snails (Gow et al. 2005). One human study, which focused on dinucleotide repeats (Huang et al. 2002,) has even found multi-step mutational events to be in the majority (63%). What accounts for this difference in the proportion of single- to multi-step mutations at different loci and in different species remains unclear, but it seem probable that some loci or alleles might be more prone to multi-step mutations than others.

The second model of mutation consists of either recombination or unequal crossing over, each of which can lead to large scale contractions and expansions in the repeat array (Richard and Paques 2000). However, there are substantial doubts that recombination acts to majorly influence microsatellite variability as it does for minisatellites (Stephan and Cho 1994; Ellegren 2004). In addition, a recent haplotype analysis (Klintschar et al. 2004) based on a total of 4900 parent-child allele transfers in 150 paternity cases found no evidence of interchromosomal exchange, but rather supported slippage as the mutational mechanism acting at microsatellite loci. Despite the absence of indisputable proof and since the strand-slippage model fits to observed mutation patterns, it is reasonable here to consider slippage as the major mechanism constantly affecting microsatellite length variability. Furthermore, on top of the forward-backward slippage process, base substitutions and short indels interrupt the repeat array and add to the polymorphism of microsatellites (see below).

Nature of microsatellites. Mutation rates may vary greatly among loci and/or between alleles of the same locus depending on the structure of the microsatellite itself (Figure 1.2). First, mutation rate increases in a given locus as the number of perfect repeats extends, i.e. long uninterrupted loci mutate more often than short loci do, and are therefore more polymorphic (Ellegren 2000). This trend has been recurrently observed in vitro (Shinde et al. 2003) as in most eukaryotes, e.g. yeast (Wierdl et al. 1997), Neurospora (Dettman and Taylor 2004), fruit flies (Schug et al. 1998; Bachmann et al. 2004), plants (Azaiez et al. 2006), barn swallows (Primmer et al. 1998) and humans (Jin et al. 1996; Brinkmann et al. 1998; Ellegren 2000; Xu et al. 2000; Huang et al. 2002; Leopoldino and Pena 2002; Webster et al. 2002; Kelkar et al. 2008). The tempo of mutation rate in relation with size remains a matter of debate, but recent studies on humans show that there might be a power or exponential relationship, rather than a linear one (Brinkmann et al. 1998; Leopoldino and Pena 2002; Lai and Sun 2003; Whittaker et al. 2003; Kelkar et al. 2008). This length-dependent trend suggests an increased instability of the replication machinery at longer repeat arrays (Wierdl et al. 1997; Kelkar et al. 2008) or simply that a long array increases the odds of slippage events compared to a shorter array.

Also in relation with allele size is the difference between the rates of motif addition and deletion resulting from the forward-backward model of slippage. In humans at least, while the rate of expansion is apparently linear and constant regardless of the microsatellite size (but see Huang et al. 2002), the rate of contraction is initially low but increases exponentially as length increases (Xu et al. 2000). This length-dependent relationship produces in the lower range allele sizes an overall increase in repeat number (Figure 1.2B,C), which results in the mutational bias observed towards repeat expansion (Rubinsztein et al. 1995; Amos and Rubinstzein 1996; Primmer et al. 1996; Cooper et al.

1999; Zhu et al. 2000; Neff and Gross 2001; Vigouroux et al. 2003). Positive directionality in short alleles could arise from a tendency for loops to occur on the elongating DNA strand and not on the template strand, presumably because the former is preferentially repaired by the mismatch mutation repair (MMR) system (Sia et al. 1997; Pavlov et al. 2003). However, the opposite directional trend has been observed in clonal snails (Weetman et al. 2002), and also in prokaryotes (Metzgar et al. 2002) and *in vitro* using *Taq* polymerase (Shinde et al. 2003). Some authors argued that when all mutations from their database are considered together, there is no clear evidence of a directional bias (Brinkmann et al. 1998; Sturzeneker et al. 2000; Xu et al. 2000; Leopoldino and Pena 2002). Alleles should rather be classified in relation to their length in order to differentiate the behaviour of short and long arrays, otherwise the differential bias in direction would go unnoticed as both processes cancel each other out (Huang et al. 2002). Indeed, longer alleles show a propensity to contract (Xu et al. 2000) (see 'Midlife crisis').

Motif length also appears to alter the rate of mutation in microsatellites (Chakraborty et al. 1997). Despite substantial interlocus variation, the mutation rate of a given array length class appears to be inversely related to motif size, i.e. dinucleotide repeats have the highest mutation rate, followed by tri- and tetranucleotide repeats (Chakraborty et al. 1997; Kruglyak et al. 1998; Schug et al. 1998; Lee et al. 1999). Few studies have included a comparison of mono-, penta- and hexanucleotide repeats with other repeat types, but there is some indication that mononucleotide microsatellites are more mutable in cultured mammalian cells (Boyer et al. 2002), in prokaryotes (Eckert and Yan 2000) and in human (Kelkar et al. 2008), and that, paradoxically, pentanucleotide repeats are less stable in yeast (Sia et al. 1997). Unfortunately, the causes underlying these differences are unknown.

The nucleotide composition of motifs adds to the interlocus heterogeneity of mutation rates (Bachtrog et al. 2000). Until recently, investigation of a motif composition effect in a given motif and array size class was beyond the scope of studies in microsatellite evolution (but see Xu et al. 2005), mainly because a great amount of data is needed in order to tease out such subtleties (Kruglyak et al. 2000). Nevertheless we know that G_{17} mononucleotide tracts are more unstable than A_{17} tracts in mammalian cell cultures (Boyer et al. 2002), but a recent analysis of microsatellite mutability in human and chimpanzee showed that this trend was reversed for shorter alleles (Kelkar et al. 2008). $(AC)_n$ and $(AT)_n$ repeats are particularly unstable when compared to other dinucleotide repeats in fruit flies (Bachtrog et al. 2000) and in human (Kelkar et al. 2008), respectively. In addition, the $(AAT)_n$ family has a higher slippage rate than four other families of trinucleotide repeats in yeast (Kruglyak et al. 2000), but the $(AAG)_n$ family has the highest mutability in human (Kelkar et al. 2008). Furthermore, the nucleotide composition of disease-associated trinucleotide repeats (CAG•CTG, CGG•CCG, and GAA•TTC) allow the formation of very stable intra- or interstrand secondary structures that are likely to induce large expansions with pathological effects (Pearson et al. 2005; Kovtun and McMurray 2008).

Finally yet importantly, the internal architecture of a microsatellite, i.e. if it is simple, compound and/or interrupted (Table 1.1), is an additional interallelic factor. Unfortunately, controversy exists between studies around the definition of a microsatellite locus, including its repetitive structure (Chambers and MacAvoy 2000), making it difficult to evaluate and compare alleged results. For practical reasons, most studies on microsatellite evolution have focused on perfect repeats, but such repeats are clearly not representative of all microsatellites in eukaryotic genomes (Almeida and Penha-Goncalves 2004). Compound microsatellites may account for 10% of all microsatellites and their variation may be more

intricate than pure repeats (Bull et al. 1999). Available information for human Y-chromosomal microsatellites suggests that (i) mutability is higher in a compound microsatellite (e.g. $[GATA]_8[GACA]_4$) than in a pure microsatellite having the same length as the longest homogeneous run in the compound locus (e.g. $[GATA]_8$), and (ii) the number of additional repeats outside the longest homogeneous array increases the variability of the latter (Kayser et al. 2004). The Y-chromosome work aside, the evolution of complex microsatellites has mainly been investigated through comparative analysis (Zhu et al. 2000). Although very interesting and informative, these studies have often provided a one-case scenario that is difficult to extrapolate to other complex loci.

Table 1.1: Classification of microsatellites relative to their basic structure.

Class	Number of repeat motif	Examples
Simple	1	$-(CA)_{12}-$
Interrupted simple	1	$-(CA)_8-(CT)-(CA)_3-$
Compound*	>1	$-(CA)_9-(GAA)_5-$
Interrupted compound*	>1	$-(CA)_9-(CAA)-(GAA)_4-$

Interruptions are critical to the evolution of all classes of microsatellites, affecting simple and compound loci over time. Imperfections occur within repeats, mainly from substitutions but also from short indels, and preferentially at the end of the array (Brohede and Ellegren 1999; Colson and Goldstein 1999). However, replication slippage can still restore the initial state by removing an interruption that is included in the loop (Harr et al. 2000). Despite a considerable theoretical debate (Kruglyak et al. 1998; Sibly et al. 2003), stability induced by interruptions within the array is well documented, e.g. in yeast (Petes et al. 1997; Rolfsmeier and Lahue 2000), in fruit flies (Goldstein and Clark 1995), in humans (Jin et al. 1996; Sainudiin et al. 2004) or mismatch-deficient cell lineages (Bacon

* Intricate forms of compound microsatellites are also sometimes dubbed complex microsatellites, e.g. Domingo-Roura X et al. (2005) Phylogenetic inference and comparative evolution of a complex microsatellite and its flanking regions in carnivores. Genetical Research 85 (3): 223-233.

et al. 2000; Boyer et al. 2008). Allele length variability is still observed in interrupted microsatellites, but generally to a lesser extent than for perfect repeats. Increased stability in interrupted arrays certainly follows from inhibition of loop formation during slippage, especially if interruptions are located closely to origins of replication (Rolfsmeier and Lahue 2000). An alternative view is that an interruption will break an array in two smaller arrays with lower intrinsic expansion rates (Boyer et al. 2008). The sequence of the interrupting base(s) has also been suggested to determine the magnitude of the effect on mutation rate (Boyer et al. 2008). Interestingly, long perfect dinucleotide repeats have been found to be typically abundant in vertebrates compared to invertebrates (Almeida and Penha-Goncalves 2004). The authors of this study assumed that interruptions accumulate with time in a repeat array (Figure 1.2D) and invoked the hypothesis of a late acquisition of long perfect dinucleotide repeats in chordate evolution. This hypothesis is concordant with the concept of a life cycle. In vertebrate genomes, long, perfect dinucleotides would then be comparatively young, still in the phase of growth and potentially close to encountering interruptions and degeneration (see 'Midlife crisis').

Genomic context. The position of a microsatellite in the genome may also influence mutational processes, contributing to an interlocus variability (Figure 1.3). Unfortunately, this area of research has not yet received enough attention. Two or more neighbouring microsatellites may well influence each other's evolution (Udupa et al. 2004), but because mutation rate varies within genomes (Ellegren et al. 2003), the mutability of microsatellites will greatly depend on the genomic composition of their flanking sequences. Factors such as G+C content and vicinity to CpG islands (Brock et al. 1999), sequence divergence (Santibáñez-Koref et al. 2001), or whether flanking sequences are composed (or not) of gene-related DNA (Metzgar et al. 2000) are certainly of key importance. In particular,

changes of tract length in a number of microsatellite loci linked to gene regulation, transcription or protein function (reviewed in Kashi and King 2006 and Li et al. 2004) might be reciprocally influenced by selective constraints. This context-dependent stability is thought to be linked to a local variation in the efficiency of the MMR system, perhaps reflecting the influence of chromosome structure, if not again that of selective constraints (Hawk et al. 2005, see also 'Selective influences'). This could explain why some microsatellites situated in conserved regions (i.e. with low mutation rates) are retained across taxa and have an apparent life span of several million years (Ross et al. 2003). These conserved microsatellites are a boon for testing the consistency of the life cycle concept, provided that priming sites are also conserved across species (Figure 1.2E, Zhu et al. 2000). Studies of genomic context effects should be eased with an ever-increasing availability of online genomic databases together with the development of search algorithms capable of finding and locating microsatellite loci, e.g. SciRoKo (Kofler et al. 2007). However, the first of such accounts reported only little effect of surrounding genomic factors on microsatellite mutability (Kelkar et al. 2008).

Large-scale mutation (duplication, translocation, recombination, genomic infection by retroviruses) of a sequence that contains or flanks a microsatellite will modify the genomic context of the microsatellite and may change the mutability of this locus, hence the course of its life cycle. However, investigating such loci is not trouble-free in the laboratory, not only because duplicated loci that are identical in length and in sequence might remain unnoticed, but also in the case of mispriming if substitutions occur in the priming site(s) of one or more copies of the microsatellite (Kayser et al. 2004). Finally, another promising area of investigation is the effect of chromosome and especially chromatin architectures on microsatellite variability, or *vice versa* (Vogt 1990), and particularly the apparent

association between certain microsatellites and recombination hotspots (Bagshaw et al. 2008).

Individual biological context. Some investigators tried to shed light on how mutational events are affected by the general biological context experienced by an individual (Figure 1.2), including its sex (Ellegren 2000). Having more mitotic divisions than females for gamete production, males are expected to exhibit more mutations per generation (Ellegren 2007). Therefore, microsatellite loci should also show a male bias in mutations (Brinkmann et al. 1998; Primmer et al. 1998; Ellegren 2000). Following from the accumulation of germ-line divisions throughout adulthood, age is also likely to be a contributing factor to heightened mutation rate of microsatellites. Although lacking support in some studies (Brohede et al. 2004; Dupuy et al. 2004), others give evidence for a positive correlation between father age and microsatellite mutation rate (Brinkmann et al. 1998; Gusmão et al. 2005). In addition, CAG expansions associated with Huntington's disease were shown to be age-dependent and to occur in the process of removing base lesions caused by oxidative damage. In addition to age and sex, mutations in MMR genes could provoke microsatellite instability in some individuals, which could sometimes initiate certain types of cancer (reviewed in Woerner et al. 2006). Finally, while only a few studies have investigated the effect of environmental stresses and stimuli on tandem repeat mutability, e.g. radiation in wheat (Kovalchuk et al. 2003) and microclimatic changes over wild barley populations (Nevo et al. 2005), and/or for rapid adaptive evolution (Marcotte et al. 1999), it is likely that mutation rates for microsatellites are influenced by these environmental variables.

Selective influences. Taxon-specific features are the fruit of evolution and a number of them are probably ground for the observed heterogeneity of microsatellite distribution and mutation rate in eukaryotic genomes (Figure 1.3). For example, canids possess a genome-wide increase in the basal germ-line slippage mutation rate compared to other carnivores (Laidlaw et al. 2007). Mode of reproduction (i.e. sexual or clonal), metabolic rate (e.g. homeotherms/poikilotherms, Neff and Gross 2001), sociality, generation time, body size, and selective adaptation, are some of the more obvious factors that might influence microsatellite mutational dynamics at the species level. Of key importance is the efficiency of the DNA machinery, involved in the replication process itself or in the correction of replication errors (Li 2008). A functional MMR system reduces the mutation rate of microsatellites between 100- and 1000-fold. Ultimately, these proteins govern the balance between enrichment and prevention of microsatellites within genomes. In a given species, MMR proteins play a role in the mutational variability among alleles, loci and individuals (Sia et al. 1997; Harr et al. 2002), and since they are driven by selective forces, are certainly the cause of differential allele distributions between species (Sainudiin et al. 2004). Moreover, the paraphernalia of proteins involved in MMR vary in number and nature among eukaryotes (Li 2008), suggesting variability in their intrinsic efficiency. In this respect, it would be of great interest to obtain in vivo measures of efficiency of the MMR system (Sia et al. 1997; Lei et al. 2004; Gu and Li 2006) that could be comparable between species.

1.2.4 Midlife crisis

Without an upper length constraint of some sort, expansion of microsatellites in eukaryotic genomes could be perpetual; instead, most microsatellites reach a pending state around a

focal length (Figure 1.2B). With high heterogeneity observed between species and loci in the literature (Ustinova et al. 2006; Vowles and Amos 2006), it is challenging to propose and define a commonly accepted size limit for microsatellites in eukaryotic genomes. Expansion of long alleles seems to be restricted to a few tens of repeats and virtually never exceeds 50 repeats (Garza et al. 1995; Stefanini and Feldman 2000; Sibly et al. 2003; Whittaker et al. 2003; Sainudiin et al. 2004; Ustinova et al. 2006), although a few larger repeat arrays have been found, e.g. in barn swallows (Primmer et al. 1996), in honey bees (Estoup et al. 1993), or some trinucleotide repeats in mammals (Pearson et al. 2005; Clark et al. 2006). Kruglyak and co-workers (1998) proposed that microsatellite growth reaches a finite upper limit since expansion by slippage is hindered by the introduction of imperfections in the repeat array, as shown earlier (Figure 2A,D). This slippage/point mutation model was an attractive attempt to explain both the absence of infinite growth at microsatellite loci and the observed steady-state distribution of repeat lengths. In practice, accounting for base substitutions and indels that break the repeat pattern of a microsatellite requires sequencing of each allele because electrophoretic-based methods fail to determine whether two alleles identical in state (IIS) are not identical by descent (IBD), a phenomenon known as size homoplasy (reviewed in Estoup et al. 2002). This is particularly problematic as allele size constraints act on the range of allele sizes, reducing the number of possible allelic states (Stefanini and Feldman 2000), thus favouring size homoplasy. Unfortunately, the accumulation of interrupts alone is not sufficient to explain the existence of a length constraint (Sibly et al. 2003; Sainudiin et al. 2004), and slippage can also still remove interruptions (Harr et al. 2000). In fact, while the rate of interruptions seems to be constant with size within a locus, there is among the upper size range not only a decrease in the rate of expansion (Huang et al. 2002, but see Xu et al. 2000), but also an excess of contractions (Figure 1.2C), e.g. in yeast (Wierdl et al. 1997), in fruit flies (Harr

and Schlötterer 2000) and in humans (Ellegren 2000; Huang et al. 2002; Whittaker et al. 2003). The likely cause is that the rate of contractions increases exponentially with length, a notion supported at least for humans (Xu et al. 2000; Huang et al. 2002). Likelihood-based simulation on almost 400 $(AC)_n$ microsatellites analysed in 123 human pedigrees totalling 680 individuals has shown that a shift in the prominence of mutational events towards contraction occurs once microsatellites exceed 20 repeats (Whittaker et al. 2003). The same study has also challenged the previous finding that multi-step contractions were significantly more frequent than single-step events in longer alleles (Huang et al. 2002). This predisposition towards contraction for longer microsatellites may be a result of selective forces to ensure that microsatellites and therefore genomes are kept in a reasonable size range. Analysis on larger pedigrees with more microsatellite markers may be necessary to tease out whether the occurrence of large deletions in long alleles involves multi-step slipped-strand mispairing, or unequal crossing-over, which has also been proposed (Richard and Paques 2000). All in all, it is expected that selection is acting against long alleles in that they may be phenotypically unfavourable (Garza et al. 1995). This is at least known for trinucleotide repeats involved in neurological disorders (Pearson et al. 2005).

1.2.5 Shortening and death

While expansions and contractions are in equilibrium and maintain the microsatellite at a focal length, interruptions can nevertheless still occur (Figure 1.2D). Eventually the accumulation of interrupts breaks the repeat pattern and leads to a blend of unique or non-repetitive DNA sequences that includes only short segments of the original repeat array. As previously reported, large deletions occur on top of interruptions and frequently involve

multi-step events, thus accelerating the process of degeneration of long microsatellites (Figure 1.2B, Taylor et al. 1999; Harr and Schlötterer 2000; Huang et al. 2002; Yamada et al. 2002; Vowles and Amos 2006). The outcome is known as the 'death' of the microsatellite (Taylor et al. 1999). The repeat arrangement is so deteriorated and scrambled that it now satisfies the definition of the initial cryptically simple sequences from which microsatellites are believed to arise, thus completing the life cycle (see 'Birth and maturation' and Figure 1.2A). However, the extent of interruptions and deletions must be dramatic to lead a microsatellite to death, primarily because shortened arrays could possibly experience a re-growth (the life cycle would then be truncated). Nevertheless, there is undeniable support that long loci accumulate interruptions and turn monomorphic (Taylor et al. 1999), so we assume that death is a slow, possibly multiphasic, process. Thus, the microsatellite death rate is very likely to be much lower than the birth rate in eukaryotes, which perfectly explains the apparent enrichment of microsatellites in eukaryotic genomes.

1.2.6 Renaissance

If the degeneration of a microsatellite locus into cryptically simple sequences can be regarded as its death, one may conceive a resurrection from cryptically simple sequences if the threshold size for expansion is reached again. Another cycle would thus start (Figure 1.2A). We acknowledge that to date no evidence is available to support this possibility, but we anticipate that with the accessibility to larger genomic databases from a variety of eukaryotic species, the death and revival of a microsatellite will soon be observed. Comparative analyses above the species level seem particularly appropriate to investigate

this issue, since an evolutionary time scale is almost certainly needed to document events of this nature.

1.3 Conservation of microsatellites in mammals

To document and understand the integrative concept of the microsatellite life cycle, which could help improve current models of microsatellite evolution, there is an implicit need to study the evolution of microsatellites above the species level. A prerequisite of such comparative studies is therefore to find microsatellite loci that are conserved in related species in order to infer their mutational history. Comparative analyses of microsatellite evolution are still fairly exceptional (e.g. Zhu et al. 2000 and Domingo-Roura et al. 2005), a shortage that could be explained by the lack of knowledge on the locations of widely conserved microsatellites and therefore stresses the need for a large-scale identification of conserved microsatellites. Such an analysis is made possible today by the accumulation of whole-genome projects, notably in mammals.

1.3.1 Comparative genomics in mammals

In recent years, the whole-genome comparative approach has upgraded from the status of pipe-dream to the status of reality. Genome sequencing projects were high on the agenda of the biological research community, with currently 82 eukaryotic genomes completed and 908 partially sequenced[*]. Many more will follow with the lowering costs of sequencing and the ever-growing interest of the scientific community.

[*] Source: http://www.genomesonline.org/, accessed 15/02/08.

Ultimately, the identification of genomic elements that have been conserved over time highlights those parts of the genome that are arguably essential to much of life (Ahituv et al. 2007; The ENCODE Project Consortium 2007). The full potential of these findings is palpable when they lead to the discovery of new genes (Pennacchio et al. 2001; Coghlan and Durbin 2007) and regulatory elements (Woolfe et al. 2005; King et al. 2007; Mrowka et al. 2007), but there might be many more outcomes from such comparative studies. It is thus important to develop comparative genomics methods to find sequences conserved between genomes; methods that could be reproduced when new genomes become available.

1.3.2 Finding conserved microsatellites

Generally regarded as neutrally evolving and highly versatile sequences, microsatellites are not expected to be retained above the species level, even more so when evolutionary distance increases. However, fuelled by the prospects of cutting the prohibitive costs of *de novo* microsatellite isolation for each species, and the possibility of comparative gene mapping (Sun and Kirkpatrick 1996), various accounts of microsatellite conservation have filled the literature. The scale of conservation reported spanned from closely related species (Schlötterer et al. 1991; Slate et al. 1998; Clisson et al. 2000; Guillemaud et al. 2000; Gonzalez-Martinez et al. 2004) to species that diverged 100+ million years ago (FitzSimmons et al. 1995; Stallings 1995; Rico et al. 1996; Ezenwa et al. 1998; Moore et al. 1998). However remarkable, these anecdotal findings were limited to a single or few loci, making it difficult to generalize to other microsatellites.

Before the turn of the last century and the advent of high-throughput sequencing, methodologies used to find conserved microsatellites were slight variations on a same theme (Schlötterer et al. 1991): the retention of a microsatellite locus that was originally isolated from a focal species, and for which primers have been specifically designed, was tested in related species using the same primer pair to amplify the hypothetical orthologous sequence. The conservation of a microsatellite in two or more species implied (1) high similarity of its flanking sequences, or at least of the primer sequences, and (2) the clearly identifiable presence of a repeat structure meeting the definition of a microsatellite. However, with increasing evolutionary distance, substitutions in the priming sites may accumulate, decreasing the odds of cross-species amplification success (Barbara et al. 2007). The lack of comparative sequence data in related taxa imply that investigators work blind when attempting to transfer microsatellite markers between related species, and generally cannot reliably design the most conserved primer pairs possible at any given locus.

1.4 Aims of the study

The general aim of this thesis has been to study the conservation and evolution of microsatellites in mammalian genomes. More specifically to use a blend of bioinformatic and comparative approaches to explore:

- The extent and patterns of conservation of human microsatellites
- The possibility to develop comparative primers in mammals
- Microsatellite evolution in mammals above the species level.

Chapter 2

2 Evolutionary conservation of human microsatellites in 16 vertebrate genomes

2.1 Abstract

The near or full completion of many vertebrate genomes, and their alignment against one another, offer a definitive approach to find conserved genomic elements. Genes and *cis*-regulatory regions are typically sought after and expected to be under selective constraints that promote their retention in related organisms. Few studies have had a focus on the genomewide conservation of genomic elements that are generally assumed to evolve neutrally. Microsatellites, a class of highly polymorphic repetitive sequences, are mostly considered to be junk DNA and too labile to be maintained in genomes over large evolutionary scale, but this view has rarely been challenged. We used the multiple-alignment of the human genome against those of 11 mammalian and five non-mammalian vertebrates to identify and examine the extent of conservation of human microsatellites. Out of 696,016 microsatellites found in human sequences, 85.39% were conserved in at least one species and 28.65% in at least one non-primate species. An overall exponential decline with increasing evolutionary time, a comparable distribution of conserved *vs.* non-conserved microsatellites in the human genome, and a positive correlation between microsatellite conservation and overall sequence conservation suggested that most microsatellites are subject to random genetic drift and are only maintained in genomes by chance, although exceptionally conserved microsatellites were also identified. Overall, A+T-rich microsatellites were the most abundant class of microsatellites, especially in non-exonic human sequences, but they also disappeared more rapidly than G+C-rich and exonic microsatellites. Comprehensive knowledge on human conserved microsatellites will prove essential to single out putative functional loci that are actively selected for, study microsatellite evolution above the species level, and help develop comparative primers useful in cross-species population genetics or comparative mapping.

2.2 Introduction

In recent years, genome sequence projects have increased in number and evolutionary scope. Despite this growing amount of comparative sequencing data available for analysis, we still have an incomplete understanding of the organization, evolution and functional landscape of eukaryotic genomes, as was emphasised by the recent findings of the ENCODE Project Consortium (Gerstein et al. 2007; King et al. 2007; The ENCODE Project Consortium 2007; Thurman et al. 2007). New approaches and ideas are continuously being developed to find sense not only in the protein-coding portions of genomes, but also in the largely more unknown non-coding regions. In particular, methods in comparative genomics (reviewed in Miller et al. 2004) have helped infer historical relationships among homologous sequences and species (Nikolaev et al. 2007; Wildman et al. 2007), estimate sequence divergence and selective constraint (Margulies et al. 2007), and predict evolutionarily conserved and/or functional sequences (Margulies and Birney 2008).

It is now well known that ~5% of the human genome is under active selection (Waterston et al. 2002; Cooper et al. 2005). Human protein-coding sequences (cds) and untranslated regions (UTRs) are found to be strongly conserved across vertebrate species (Roest Crollius et al. 2000), but they cover only ~2 % of the human genome (International Human Genome Sequencing Consortium 2004). Pairwise and multiple genome comparisons demonstrated that ~3.5% of the non-coding DNA sequence, introns and intergenic regions (IGRs), is under negative selection across the Mammalia (Waterston et al. 2002; Lindblad-Toh et al. 2005; Mikkelsen et al. 2007), a figure that could be greatly underestimated if the neutral rate of evolution in reference sequences is also underestimated (Pheasant and

Mattick 2007). A substantial amount of non-coding sequence is also shared with more distant vertebrates, including avian, amphibian and fish species (Siepel et al. 2005; Venkatesh et al. 2006; Loots and Ovcharenko 2007).

The method of choice to identify human conserved elements at the genome scale relies on sequence alignments and model prediction of constrained segments (Cooper et al. 2005; Siepel et al. 2005; Prabhakar et al. 2006). More specific genomewide approaches have also been used to narrow the search down to clearly identifiable classes of genomic elements, such as short regulatory motifs (Xie et al. 2007), transposable elements (Lowe et al. 2007) and microsatellites (simple repeats). However, with the exception of the recent analysis of the platypus genome (Warren et al. 2008, Chapter 3), studies of genome-scale microsatellite conservation in vertebrates were typically limited to pairwise comparisons of closely related species (Kayser et al. 2004; Vowles and Amos 2006; Kelkar et al. 2008), and thus lacked the depth needed to implement fully the comparative method and estimate the true extent of microsatellite conservation.

Microsatellites are arrays of short, tandemly repeated, DNA motifs (1-6 bp) found throughout the genomes of both prokaryotes and eukaryotes (Buschiazzo and Gemmell 2006). Their distribution and density in genomes appear to be non-random but can vary greatly, even between closely related species (Tóth et al. 2000; Warren et al. 2008). Microsatellites have gained notoriety in medical genetics with evidence of association with colorectal, endometrial, and various other cancers (Woerner et al. 2006), and the implication of unstable repeats in ~30 human hereditary disorders (Mirkin 2007). Other microsatellites, in contrast, are thought to play an advantageous role in evolution (Kashi and King 2006). However, microsatellites have attracted the widest interest as genetic

markers for population genetics, gene mapping, forensics or paternal investigation (Schlötterer 2004).

Traditionally regarded as neutrally evolving and highly polymorphic sequences, microsatellites are not expected to be retained in different species, particularly when evolutionary distance increases (Schlötterer et al. 1991; Barbara et al. 2007). However, these assumptions are questioned by theoretical expectations (Tachida and Iizuka 1992; Stephan and Kim 1998) and numerous observations of microsatellite conservation, not only in closely related species (Schlötterer et al. 1991; Blanquer-Maumont and Crouauroy 1995; Primmer et al. 1996; Gemmell et al. 1997; Crawford et al. 1998; Slate et al. 1998; Guillemaud et al. 2000; Gonzalez-Martinez et al. 2004), but also in species that diverged 100+ million years ago (FitzSimmons et al. 1995; Rico et al. 1996; Ezenwa et al. 1998; Moore et al. 1998). These efforts were mainly fuelled by the prospects of transferring microsatellite markers between related species to promote comparative gene mapping (Sun and Kirkpatrick 1996), cutting the development costs of *de novo* microsatellites (Barbara et al. 2007), and the opportunity to study microsatellite evolution above the species level (Zhu et al. 2000). Unfortunately, reports of microsatellite conservation were limited to one or few loci, and genomewide searches for homologous human microsatellites were limited to comparisons with chimpanzee (Kayser et al. 2006; Vowles and Amos 2006; Kelkar et al. 2008) or other close primate relatives (Raveendran et al. 2006).

In response to the significant lack of evolutionary scope in previous analyses of microsatellite conservation, it is timely to develop a reliable method to identify, at the genome scale, human microsatellites conserved in mammals and beyond. In this study, we used the publicly available multiple alignment of the human genome against the genomes

of 16 vertebrates, including 11 mammalian species, to investigate the extent and patterns of conservation of human microsatellites. In particular, we suggest that most microsatellites are maintained by chance, although there might be exceptional cases of broadly conserved microsatellites that have been selected for or that lie in regions of the genome under strong selection. We discuss the implications and applications of conserved microsatellites in evolutionary genomics and genetics, e.g. cross-species transferability of microsatellite markers for population-based analyses.

2.3 Materials and Methods

2.3.1 Vertebrate sequences

The 17-way vertebrate alignments (17-WA) available on the UCSC Genome Browser for each human chromosome were downloaded by anonymous FTP from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz17way/. MAF-formatted blocks were extracted and converted to FASTA format using a stand-alone version of Galaxy (Giardine et al. 2005) downloaded from http://main.g2.bx.psu.edu/. Due to the large size of the alignments for chromosomes 1 to 4, the files were split in half; this had no consequence except that an additional step to merge results for each respective chromosome was required. Sequence gaps were removed using the degapseq module from the EMBOSS 5.0 package (Rice et al. 2000).

2.3.2 Microsatellite search and classification

Perfect and imperfect microsatellites (motif length: 1 to 6 bp) were searched in ungapped sequences using SciRoKo 3.1 (Kofler et al. 2007) with fixed penalty parameters (score: 12, mismatch penalty: 4, SSR seed min. length: 3, SSR seed min. repeats: 3, max. mismatches at once: 3). Genomic intervals of microsatellites in each vertebrate genome were recorded with block number, standardized repeat motif (Kofler et al. 2007), array length, and number of imperfections. Microsatellites in the alignment of the Y chromosome were not included in analyses unless stated in the text. Human microsatellites lying in segmental duplications >1 kb and >90% identity (Bailey et al. 2001), and non-human overlapping intervals (5 bp minimum cut-off) were removed. Intervals overlapping with repeats other than simple repeats or low complexity sequence (Smit et al. 1996-2007) were also discarded. Segmental duplication and repeat data were retrieved from the UCSC Table Browser (Karolchik et al. 2003). If the sequence 25 bp upstream and downstream of a microsatellite interval did not contain another microsatellite, the microsatellite was classified as 'simple'. Microsatellite segments were merged and classified as 'compound' if they were 5 bp or less apart from each other or were overlapping by 5 bp or less, 'linked' if they were separated by 5 to 25 bp, or 'mixed' if they contained both linked and compound portions. This series of operations produced, in human, a dataset of 696,016 microsatellites covering 19.5 Mb of the human genome (0.70% of human sequences in the 17-WA).

2.3.3 Microsatellite conservation

Positions of non-human microsatellites were converted to the hg18 human assembly using the liftOver utility and chain files (Kent et al. 2003) available at the UCSC Genome Browser (Karolchik et al. 2003). Converted intervals overlapping with human repeats other than simple or low complexity repeats were discarded. The fraction of human microsatellites overlapping with any of the converted microsatellite positions indicated conserved sites. We found 594,340 human microsatellites conserved in at least one species, i.e. 85.0 % of the initial dataset.

2.3.4 G+C composition

We classified microsatellites according to the G+C composition of their standardized motif as given in SciRoKo's output (Kofler et al. 2007). G+C-rich motifs had a majority of strong nucleotides (G or C), whereas A+T-rich motifs contained a majority of weak nucleotides (A or T). ATGC-eq motifs had a composition of strong and weak nucleotides at equilibrium. For practicality, repeat segments forming compound, complex and linked microsatellites were treated as individual microsatellites.

2.3.5 Genomic location

A tentative canonical list of 17,260 non-overlapping human nuclear genes was produced from the UCSC Genome Browser and used to locate human microsatellites conserved in coding exons, 3'-UTRs, 5'-UTRs, introns or intergenic regions (IGRs). Conserved

microsatellites spanning more than one element were positioned in the element with the longest overlap. When an equal overlap existed, we positioned the microsatellite following the preferential order given above.

2.3.6 Statistical analyses

Genomic features were based on annotations of human autosomes obtained from the UCSC Genome Browser and were calculated in 1 Mb windows using Galaxy. Densities of microsatellites were based on sequence length excluding segmental duplications and repeats, unless stated otherwise. Windows with low sequence coverage and high content of repeats and segmental duplications were excluded (i.e. windows with >70% of their length annotated as gaps and segmental duplications, and windows with >90% of their length annotated as gaps, segmental duplications and repeats). Again, these repeats do not include low complexity or simple repeats. This treatment excluded 233 windows out of 2857. We considered smaller window sizes (500 kb and 250 kb), but selected 1 Mb windows as only a negligible number of these contained no microsatellite conserved in at least 3 non-primate species (23 out of 2624 windows). Spearman's rank-order correlation tests were performed using the R package (www.r-project.org).

2.3.7 Method assessment

We sought to assess the reliability of our method to find conserved microsatellites using whole-genome alignments. Our approach essentially relies on the quality of the UCSC multiple alignments, therefore to assess the validity of the identified conserved

microsatellites, we compared their positions with regions previously found to be suspiciously aligned in the 17-WA of human chromosome 1 using a statistical assessment (Prakash and Tompa 2007). Any conserved microsatellite overlapping with regions identified as suspiciously aligned may be considered suspicious too.

2.4 Results

2.4.1 Microsatellites in the alignment

Using current algorithms, microsatellite sequences do not align well and, at first sight, might resemble sequences with no common ancestry, so the statistical approach generally applied to identify conserved regions from multiple alignments, which assumes a 'perfect' alignment (Margulies and Birney 2008), is inappropriate for microsatellites. An alternative approach, applied in recent studies of human-chimpanzee comparisons (Kayser et al. 2006; Vowles and Amos 2006; Kelkar et al. 2008), is to identify all microsatellites in each genome and find homologies by comparing positions in a pairwise alignment. Although efficient, this task may become impractical when many genomes are compared, and probably disproportionate when dealing with highly divergent species that are not expected to share many microsatellite sequences. To circumvent the additional computational time and space that this full-genome approach would require, we sought to narrow our investigation down to a subset of genomic sequences already aligned to each other, and thus very likely to contain the subset of genomic microsatellites that are conserved. The publicly available alignment of the human genome against 16 vertebrate genomes, viz. 17-way alignment (17-WA), provided a timely framework for our analysis. It includes 2

finished genomes, 11 high quality genomes (4-7.9X) and 4 low-coverage (2X) genomes (Table 2.1).

As expected, there was a negative relationship between the size of sequence aligned to the human genome (alignability) and both the phylogenetic distance from human to each comparison species and times of divergence from the common ancestor (Table 2.1). We nevertheless found large differences in alignability between species whose ancestors share a common ancestry with human. For example, the relatively low amount of human sequence aligned to the mouse genome (~37%) compared to the dog and cow genome (~57% and 51%) reflects a particularly high rate of sequence evolution and large deletions occurring in the rodent lineage (Waterston et al. 2002; Lindblad-Toh et al. 2005), and is also well known from studies of karyotype evolution (Ferguson-Smith and Trifonov 2007).

To construct our initial dataset of microsatellites in human sequences, we used SciRoKo 3.1 with fixed penalty parameters. Results of microsatellite mining in genomes are dependent on how one decides to define a microsatellite, and on the software and built-in options one chooses to automate this non-exhaustive task (Leclercq et al. 2007; Sharma et al. 2007). Our approach aimed at (i) using a fast, flexible, reproducible and user-friendly program, and (ii) finding perfect and imperfect microsatellites, with a repeating motif of size 1-6 bp, and no shorter than 12 bp/3 perfect repeats. While tolerating the identification of rather short arrays, which could help document the concept of microsatellite life cycle (Buschiazzo and Gemmell 2006), our search parameters were purposely conservative regarding purity: imperfect microsatellites that maintained a clear repeat pattern were included, but low complexity DNA and over-degenerated repeat sequences were ignored, thus avoiding the need for additional filtering of spurious sequences.

Table 2.1: Species in the 17-way alignment (17-WA). Cov: genome sequence coverage; Dist 17-WA: distance to human, branch lengths retrieved from 17-way UCSC data; Div: divergence time (Myr ago), retrieved from TimeTree (http://www.timetree.org/), except times in italics (Bininda-Emonds et al. 2007); Human align: aligned fraction of the human genome.

Scientific name	Common name	UCSC ID	Cov	Dist 17-WA	Div	Size sequenced (Gb)	In 17-WA (including overlaps)	Human align
Homo Sapiens	Human	hg18	Fin	0	0	2.86	2.79	100%
Pan troglodytes	Chimp	panTro1	4.0x	0.014261	5.8	2.73	2.67	95.59%
Macaca mulatta	Rhesus	rheMac2	5.1x	0.090162	33.3	2.87	2.49	89.39%
Mus musculus	Mouse	mm8	Fin	0.467712	*91.8*	2.60	1.04	37.19%
Rattus norvegicus	Rat	rn4	7.0x	0.472423	*91.8*	2.57	0.98	35.39%
Oryctolagus cuniculus	Rabbit	oryCun1	2.0x	0.368189	*91.8*	2.08	1.01	36.31%
Canis familiaris	Dog	canFam2	7.6x	0.265129	*98.9*	2.40	1.59	56.95%
Bos Taurus	Cow	bosTau2	6.3x	0.27658	*98.9*	2.62	1.42	51.01%
Dasypus novemcitus	Armadillo	dasNov1	2.0x	0.256315	*101.1*	2.15	0.97	34.62%
Loxodonta africana	Elephant	loxAfr1	2.0x	0.267708	*101.3*	2.30	1.04	37.39%
Echinops telfairi	Tenrec	echTel1	2.0x	0.422614	*101.3*	2.11	0.72	25.90%
Monodelphis domestica	Opossum	monDom4	6.5x	0.71192	*147.7*	3.50	0.37	13.38%
Gallus Gallus	Chicken	galGal2	6.6x	0.984662	323.6	1.05	0.11	3.76%
Xenopus tropicalis	Frog	xenTro1	7.4x	1.435721	360.0	1.63	0.06	2.34%
Danio Rerio	Zebrafish	danRer3	6.5-7x	1.747963	450	1.63	0.06	2.24%
Takifugu rubripes	Fugu	fr1	5.7x	1.698263	450	0.32	0.05	1.77%
Tetraodon nigroviridis	Tetraodon	tetNig1	7.9x	1.65775	450	0.35	0.06	1.96%

When microsatellites in segmental duplications (Bailey et al. 2001) and repeats (Smit et al. 1996-2007) were discarded to restrict our analysis to orthologous microsatellites, we obtained a total of 696,016 human microsatellites in autosomes and the X chromosome (Appendix, Table 1). We classified microsatellites as simple, compound, linked and mixed (see Methods), not only because clustered microsatellites tend to evolve differently (Buschiazzo and Gemmell 2006), but also to ensure that two neighboring microsatellites

were separated by at least 25 bp of 'unique' sequence, a sufficient length to design a potential primer for comparative PCR-based analysis (Chapter 4). Our dataset was thus represented by 89.15% simple, 6.32% compound and 3.51% linked human microsatellites, and simple microsatellites were represented by 19.28% mono-, 28.36% di-, 15.63% tri-, 28.16% tetra-, 6.94% penta- and 1.63% hexanucleotide repeats (Appendix, Table 1).

Similar datasets of microsatellites were constructed in all other genomes (Appendix, Table 1), except that microsatellites in intragenomic segmental duplications were not identified and discarded. Instead, we removed microsatellites in overlapping genomic intervals (see Methods), indicating those sites that aligned to human duplicated segments. Indeed, every alignment block in the 17-WA represents one, and only one, human interval, but the same non-human interval can be assigned to one or more human intervals.

We first compared microsatellite abundance in every genome relative to the human genome (Figure 2.1 and Appendix, Table 1), and found proportions ranging from 87.88% in chimpanzee to 15.52% in opossum for mammals, and to 1.24% in fugu for vertebrates. These results are positively correlated with the amount of sequence aligned in each species (Sperman's rank correlation, $\rho = 0.89$, $P<0.0001$), and are strongly dependent on phylogenetic distance.

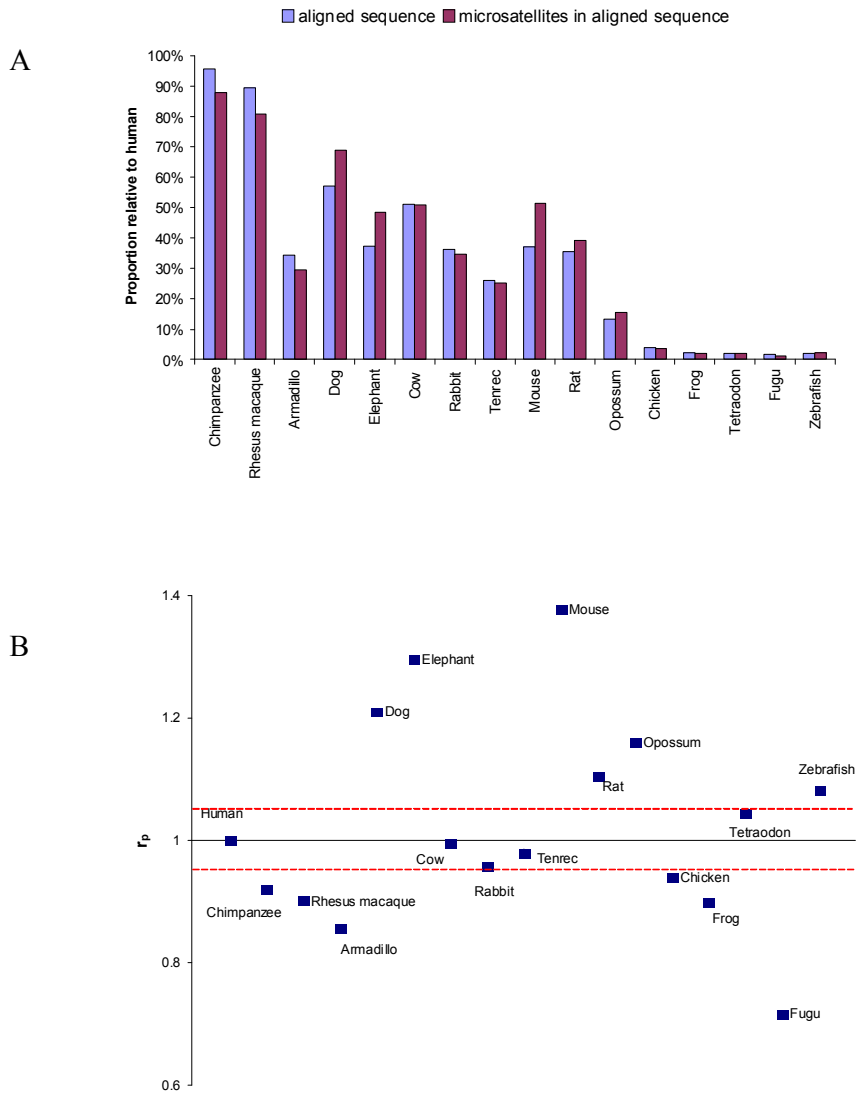


Figure 2.1: Species-specific microsatellite enrichment. (A) Alignability to the human genome, and conservation of human microsatellites in vertebrate species. **(B)** Scatter plot showing the ratio (r_p) of percentage of alignment to percentage of microsatellite conservation relative to human. Dotted lines represent a 5% significance threshold. Species are arranged from left to right by increasing distance (substitution rate) from human (Miller et al. 2007).

By measuring the ratio of the percentage of microsatellite abundance to the percentage of human sequence that aligns to every genome (Figure 2.1A and Appendix, Table 1), we also obtained an indication of whether sequences from each species were enriched or impoverished for microsatellites compared to human sequences (Figure 2.1B). Rather than

following a phylogenetic trend, this ratio demonstrated species-specific enrichment. Microsatellites were especially enriched in mouse, elephant and dog, whereas sequences from armadillo, frog, and fugu were particularly depleted in microsatellites in comparison to the human genome. These differences could be caused (i) by species-specific microsatellite birth and death events (Buschiazzo and Gemmell 2006), and/or (ii) by the species-specific nature and alignability of sequences. Our enrichment results are concordant with independent analyses of microsatellite coverage in the whole genome sequence of mouse, dog, opossum, and chicken (Waterston et al. 2002; Warren et al. 2008), thus we would favor the former hypothesis. A higher rate of microsatellite death decreases the chance of conservation, whereas a high birth rate has no direct consequence (unless births occurred in an ancestor common to two or more species).

2.4.2 Phylogenetic extent of conserved human microsatellites in vertebrate genomes

Using pairwise alignment chains between each comparison species and human (Kent et al. 2003), a strategy similar to that employed in previous works comparing microsatellites between human and chimpanzee (Kayser et al. 2006; Vowles and Amos 2006; Kelkar et al. 2008), we converted the genomic positions of all microsatellites to positions in the human assembly (UCSC hg18), and looked for overlaps. Overlapping sites indicated homologous microsatellites. We thus define conserved microsatellites as orthologous arrays of short tandem repeats found in regions that are that similar to the human genome in one or several species that they could be aligned with the BLASTZ/MULTIZ algorithms (Figure 2.2).

Of 696,016 microsatellites identified in human sequences, 594,340 (85.39%) were found to be conserved in at least one comparison species, whereas 199,403 (28.65%) were conserved in at least one non-primate species (Appendix, Table 2). The fraction of conserved human microsatellites decreased from 87.74% in chimpanzee to 1.71% in opossum for mammals, and to 0.16% in fugu for vertebrates (Appendix, Table 1). Although these results demonstrate a much higher extent of microsatellite conservation than previously found in mammals (Moore et al. 1998), they illustrate a dramatic decline of human microsatellite conservation across vertebrates, confirming the view that microsatellites are presumably-neutral sequences that are rapidly cleared from the genome by random genetic drift.

We found that microsatellite conservation decayed exponentially with increasing phylogenetic distance from human (Figure 2.3A), a pattern consistent with the neutral expectation and a recent analysis of sequence conservation in 28 vertebrate genomes (Miller et al. 2007).

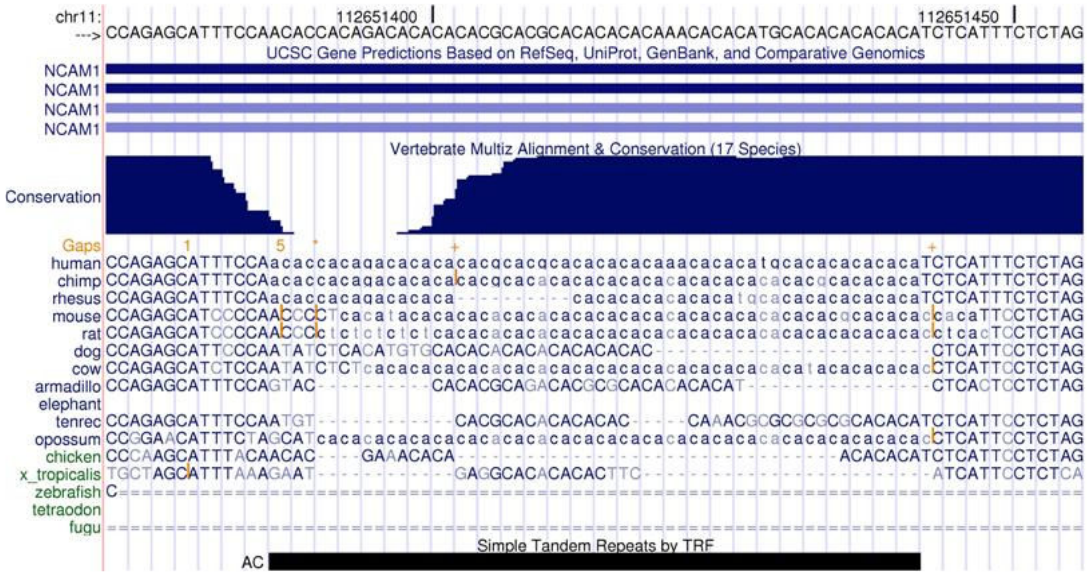
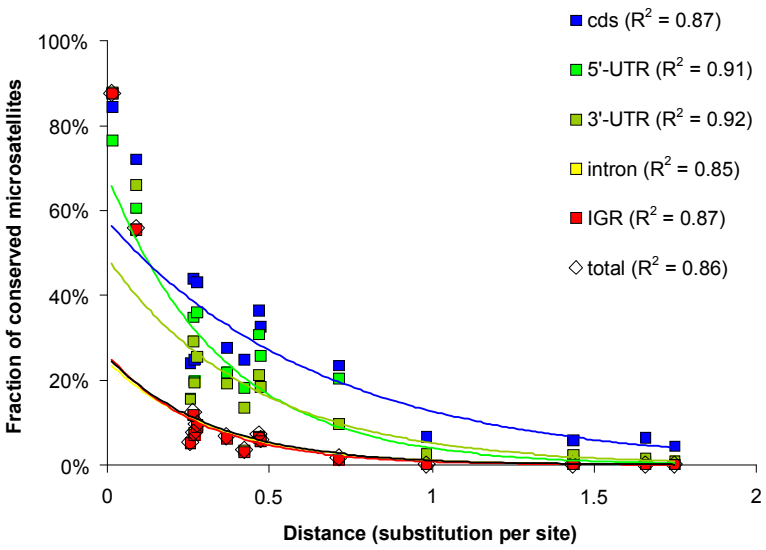


Figure 2.2: Conservation of a $(CA)_n$ microsatellite in the 3'-UTR of the human *NCAM1* gene (Moore et al. 1998). hg18.chr11:112,651,373-112,651,456.

A



B

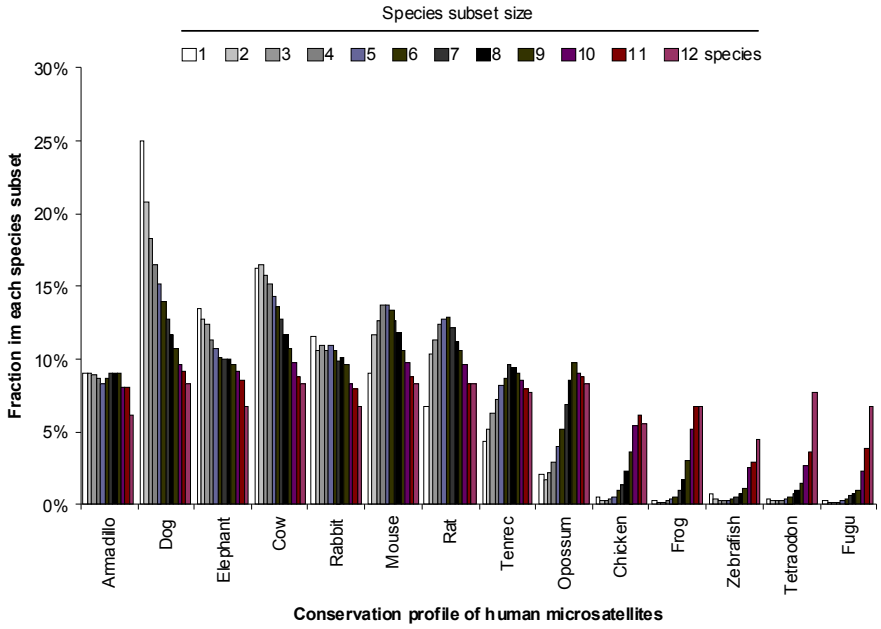


Figure 2.3: Phylogenetic extent of conservation of human microsatellites. (A) Decay of conservation in different genomic locations as a function of phylogenetic distance. The distance is measured as the total substitutions per 4D site on each of the branches connecting human to the comparison species.

(B) Conservation profile of human microsatellites in each comparison species. The range of conservation represents the range of species that share the same microsatellites with human, from exclusive (1 species) to wide (12 species). No microsatellite was found in 13 species, and only one in all 14 species. Bar plots of identical range add up to 100%. Species are arranged from left to right by increasing distance from human (Miller et al. 2007). Primates were excluded.

To explore patterns of microsatellite conservation further, we examined the proportion of human microsatellites conserved within species subsets (Figure 2.3B). Here, primates were excluded to allow defining differences among species distantly related to human. A profile skewed to the left indicates a species that share microsatellites relatively exclusively with human, whereas a profile skewed to the right indicates a species that mostly shares microsatellites that are broadly conserved. Under a neutral model of evolution, these scenarios would be typical of species that are respectively closer, e.g. dog, and more distant, e.g. zebrafish, to human. Figure 2.3B shows that this expectation is in relatively good agreement with our observations, with intermediate stages between the two extremes. In fact, only species with a mere 2X coverage did not perfectly fit with this general pattern, e.g. armadillo, the closest species to human if 4-fold degenerate (4D) site substitutions are used to measure phylogenetic distance, which revealed a flat conservation profile instead of the expected skew to the left. We believe however that profiles from 2X covered genomes are not complete and that any premature interpretation should therefore be avoided.

Overall, our results suggest that drift predominantly affected the decay and thus the conservation, of microsatellites in vertebrate genomes; a pattern consistent with the neutral expectation.

2.4.3 Interchromosomal distribution of human conserved microsatellites

We sought to investigate whether there was any difference in the distribution of genomic and conserved microsatellites at the chromosome level. Our initial dataset of human

microsatellites (MSATs) was based on the microsatellite content in the aligned and unique fraction of the human genome (see Methods). This fraction represented only ~37% of the human genome, and was fairly heterogeneous between chromosomes (Appendix, Table 2). Chromosomes 18 and 13 were highly represented in the 17-WA (42.26% and 41.72%, respectively) while X, 19, 22 and 16 had the lowest representations (28.42%, 30.14%, 33.43% and 34.17%, respectively). The origin of this disparity is essentially interchromosomal differences in (i) the amount of gaps in the sequence, mostly a result of high heterochromatin content, (ii) content of segmental duplications and repeats (SD+R), (iii) sequence alignability with other genomes (Appendix, Figure 1), and (iv) might also reflect differences in other genomic features that may affect microsatellite distribution, e.g. gene density. Only human, chimp and mouse Y chromosomes were made available in the 17-WA, hence the exceptionally low representation of the human Y sequence (13.51%) that prompted us to exclude the Y data from our analyses (Appendix, Table 2).

Acknowledging this interchromosomal heterogeneity in alignability is obviously important to appreciate the interchromosomal differences in microsatellite abundance (Figure 2.4). Based on total ungapped length of chromosomes, MSAT density appeared particularly homogeneous among autosomes (249.3±10.6 MSAT/Mb), although the X chromosome exhibited a lower density (207.1 MSAT/Mb). Conversely, when densities were based on the length of (SD+R)-free and ungapped sequence, i.e. the portion practically analyzed, the overall picture was comparably heterogeneous (495.7±37.2 MSAT/Mb): chromosomes 19, 16, 20, X and 22 showed a relative increase in MSAT density (639.18; 550.47; 512.55; 535.14 and 508.92 MSAT/Mb, respectively) while chromosomes 13 and 18 showed a slight decrease (470.54 and 474.41 MSAT/Mb, respectively). This latter density measure also compared better with chromosomal differences found in a genomewide scan of human

microsatellites (Subramanian et al. 2003), and confirmed that the MSAT dataset, which we refer to as our background distribution, represented well the overall distribution of microsatellites in the human genome.

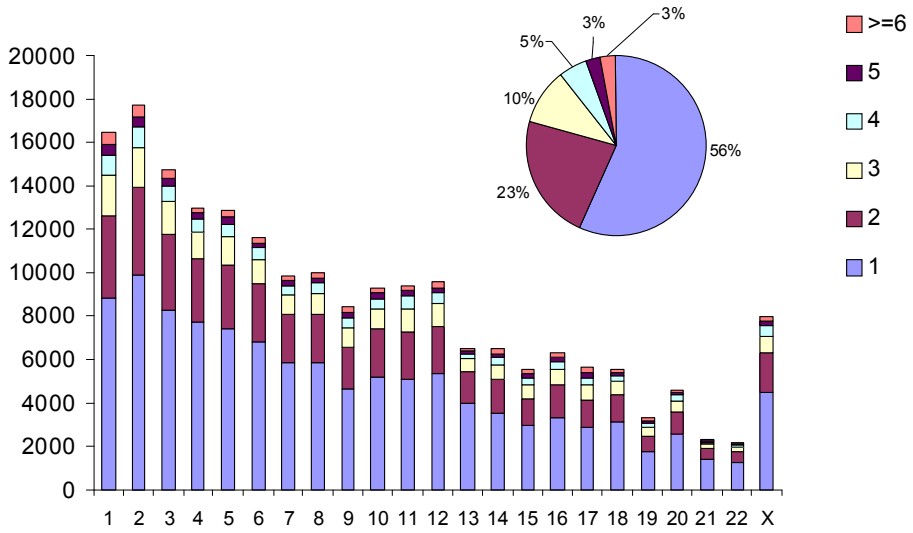


Figure 2.4: Distribution of human microsatellites conserved in non-primates species. The number of species is color-coded as indicated in the legend.

To depict any interchromosomal differences in the extent of microsatellite conservation, we counted microsatellites conserved in increasing number of species; as expected, there was a rapid decline of conserved microsatellites with increasing species number, regardless of the chromosome examined (Figure 2.4). We found it pertinent to compare MSAT abundance with numbers of microsatellites conserved in (i) at least one of all 16 species (human conserved microsatellites, HCMs), (ii) at least one of the non-primate species (NPMs) and (iii) at least three of the non-primate species (NP3Ms). At the genome scale, the three inclusive subsets represented 85.39%, 28.65% and 5.98% of the initial dataset, respectively (Appendix, Table 1 and Figure 2). At the chromosome level, proportions of HCMs were strikingly homogeneous (84.28%-86.61%) with the exception of chromosomes 19, X and 22 (77.10%, 78.86% and 82.13%). Accordingly, the alignability

of human chromosomes 19 and X was the lowest among eutherian genomes (Appendix, Figure 3). This trend was particularly true for primate genomes, in which most microsatellites contained in the HCM dataset (66.45%) were found, greatly influencing the overall distribution (Appendix, Figure 3).

When primate-specific microsatellites (PSMs) were excluded to focus on NPMs, proportions of conservation were more heterogeneous among human chromosomes (25.44%-30.68%), although chromosome 19 still showed a distinctively low proportion (22.07%). When NP3Ms only were considered, interchromosomal differences in the extent of human microsatellites were manifest (4.13%-8.09%), and did not follow previous observations, e.g. chromosome 19 had a comparatively average proportion of microsatellites conserved in at least three non-primate species (5.71%). Yet again, these results might be caused by the uneven alignability of human chromosomes to other genomes: chromosome 19 was highly represented in species distant to human, i.e. in opossum but especially in non-mammalian vertebrates (Appendix, Figure 3), and thus correlated with a relatively higher proportion of NP3M conservation. Chromosomes 1, 11, 15, 16, 17 and 22 also showed high representation in distant species and high proportion of NP3M, while chromosomes 4 and 13 showed the contrary dispositions. Strikingly, the former group had the highest gene densities in the human genome, whereas chromosomes 4 and 13 occupied the last ranks with chromosome 18 (based on our canonical list of UCSC genes). In addition, the former group contained proportionally more NPMs in exons than the latter group (Appendix, Figure 4).

Our results thus suggested that broadly conserved microsatellites tend to cluster in gene-rich chromosomes, while microsatellites with lower conservation were more evenly

distributed in the human genome. This may not be surprising as different constraints have shaped the evolution of particular genomic regions: human non-coding sequences are often under more selective constraint than exonic regions in mammalian genomes (Cooper et al. 2005), but are much less retained in non-mammalian vertebrates than exonic regions (Siepel et al. 2005). Overall, our results of interchromosomal distribution of conserved microsatellites showed that the distribution of conserved microsatellites broadly corresponded to the overall distribution of aligned, hence conserved, genomic sequences, and suggested that a finer scale analysis of microsatellite distribution in relation to other genomic elements, such as genes, would help understand why microsatellites in different chromosomes were differentially maintained in genomes.

2.4.4 Megabase distribution of human microsatellite conservation

To inspect what could drive the distribution of conserved microsatellites at a finer scale than the chromosomes level (Figure 2.5) and thus attempt to understand the causes of microsatellite conservation in genomes, we measured densities of human microsatellites (MSATs) and human microsatellites conserved in at least one species (HCMs), in primates only (PSMs), in non-primate species (NPMs) and in at least 3 non-primate species (NP3Ms) in 1Mb windows of autosomes. We carried out comparisons with sequence composition (G+C content), genomic elements (gene density and repeat coverage), and four measures of evolutionary change; two derived from the human genome (recombination rate and SNP density), and two derived from genomic comparisons (coverage in conserved, "indel-purified", intervals, *viz.* cIND, and density of conserved transcription factor binding sites, *viz*. tfbsCons). Preliminary correlation analyses between

these factors confirmed results of previous studies of the human genome (e.g. Fullerton et al. 2001 and Lander et al. 2001): G+C content covaried positively with gene density, SINE density and recombination rate, but was inversely correlated with LINE and LTR density (data not shown).

Table 2.2: Covariation between human microsatellites and other genomic features. Left to right: genomic microsatellite density, G+C content, gene density, SINE, LINE and LTR coverage, average recombination rate, indel-purified sequence coverage, SINE coverage, SNP density and density of conserved transcription factor binding sites. Source: UCSC Genome Browser. Spearman's rank correlation factor ρ, P-value significance: 0<***<0.001<**<0.01<*<0.05<not significant (n.s).

Density	MSAT	G+C	Gene	SINE	LINE	LTR	R_{recomb}	SNP	cIND	tfbs
MSAT	-	n.s.	-0.22^{***}	0.11^{***}	-0.37^{***}	-0.26^{***}	-0.33^{***}	-0.05^{**}	0.41^{***}	0.40^{***}
HCM	0.98^{***}	-0.07^{***}	-0.25^{***}	0.14^{***}	-0.32^{***}	-0.24^{***}	-0.31^{***}	-0.09^{***}	0.45^{***}	0.42^{***}
PSM	0.90^{***}	-0.11^{***}	-0.27^{**}	0.17^{***}	-0.28^{***}	-0.12^{***}	-0.26^{***}	n.s.	0.16^{***}	0.15^{***}
NPM	0.84^{***}	n.s.	-0.19^{***}	0.08^{***}	-0.27^{***}	-0.29^{***}	-0.28^{***}	-0.16^{***}	0.67^{***}	0.61^{***}
NP3M	0.63^{***}	0.17^{***}	0.04^{*}	0.15^{***}	-0.35^{***}	-0.41^{***}	0.25^{***}	-0.23^{***}	0.74^{***}	0.75^{***}
A+T-rich	-	-0.33^{***}	-0.38^{***}	0.24^{***}	-0.04^{*}	-0.14^{***}	-0.13^{***}	-0.21^{***}	0.61^{***}	0.48^{***}
G+C-rich	-	0.65^{***}	0.45^{***}	0.51^{***}	-0.62^{***}	-0.54^{***}	0.35^{***}	-0.08^{***}	0.41^{***}	0.57^{***}
ATGC-eq	-	n.s.	-0.24^{***}	0.17^{***}	-0.20^{***}	-0.17^{***}	-0.32^{***}	-0.07^{***}	0.54^{***}	0.46^{***}

First of all, we found a general positive correlation between MSAT density and densities of conserved microsatellites (Table 2.2). Figure 2.5 clearly illustrates the similar distribution between HCMs, NPMs and NP3Ms. Altogether, these findings agreed with our chromosome level analysis, and support the hypothesis that microsatellite conservation declined in mammalian genomes through random genetic drift. However, a weaker statistical significance for broadly conserved microsatellites, attributable to megabase segments containing a higher than usual proportion of NP3Ms (data not shown), indicated that a small subset of HCMs might well be maintained in genomes in a non-neutral fashion.

Table 2.2 also shows correlations between microsatellites and other genomic features. As a whole, MSAT density was negatively correlated with gene density, LINE, LTR and

recombination, and weakly associated with SINE coverage. The strongest relationship (ρ>0.40) appeared between MSATs and measures of sequence conservation (cIND and tfbsCons), but we believe that our biased alignment approach, i.e. scanning microsatellites in sequences known to possess some level of conservation, might best explain this association.

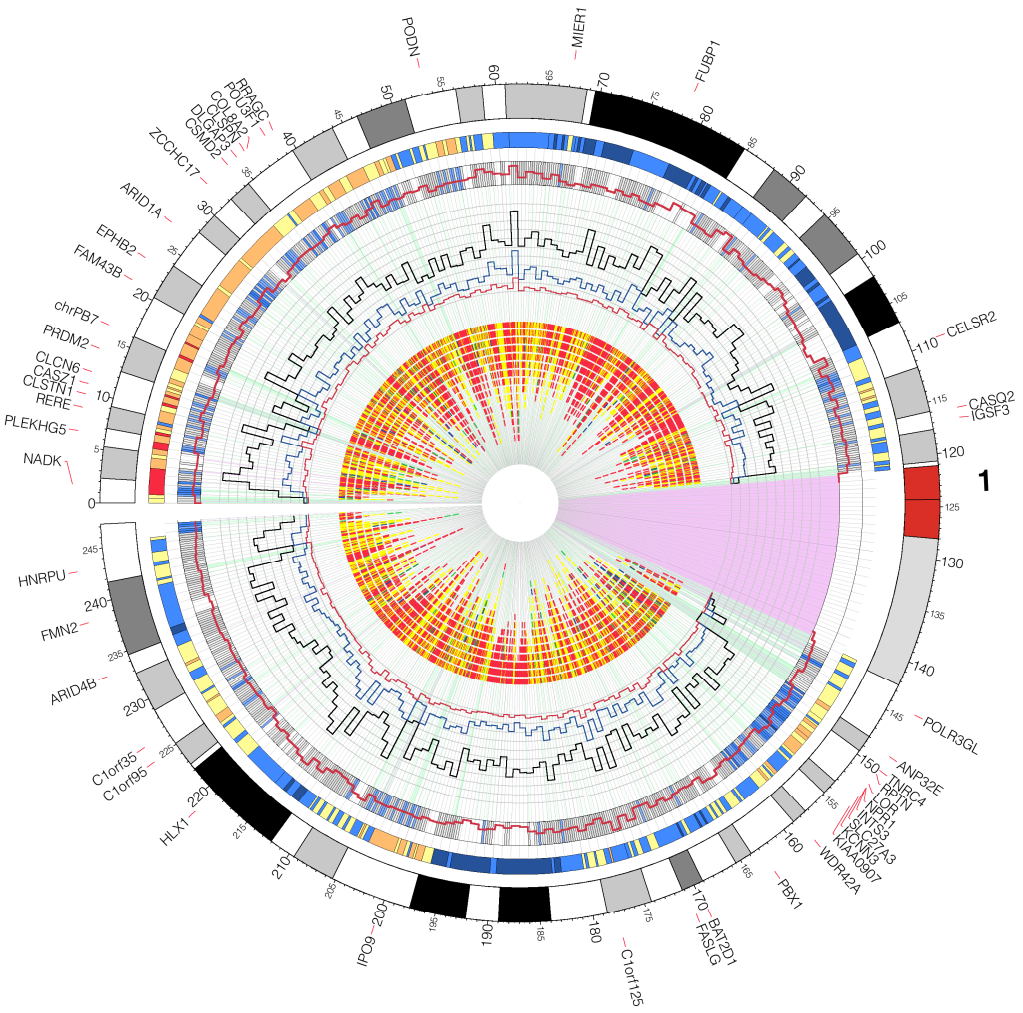


Figure 2.5: Distribution of conserved microsatellites on human chromosome 1. Inwards: names of genes containing NP3Ms (length > 21 bp) in coding exons; ideogram of chromosome with G-banding; isochore-banding (Costantini et al. 2006): red (GC>53%), orange (46-53), yellow (41-46), blue (37-41) and dark blue (<37); heatmap of gene density in 250kb-windows (increasing from white to dark blue) and histogram of conserved region (cIND) density (red); histograms of HCM density (black), NPM density (blue) and NP3M density (red) in 1Mb-windows; NPM tiles with color-coded location (IGR: red; intron: yellow; UTR: green; cds: blue); regions not represented in analysis are highlighted (segmental duplications: light green; sequence gaps in the assembly: light purple).

To explore whether G+C composition of microsatellites affected the genomic distribution of conserved microsatellites, we grouped NPMs into G+C-rich, A+T-rich and balanced-composition (ATGC-eq) microsatellites. G+C-rich NPMs were found to cluster in G+C-rich regions (Table 2.2), and were therefore typically, though weakly, associated with genes, SINEs and high recombination rate, and inversely correlated to LINE and LTR density. A+T-rich NPMs showed a contrary disposition: they were preferentially found in A+T-rich, gene-poor and low recombination rate regions, but had a relationship with repeats similar to that of G+C-rich NPMs. ATGC-eq NPMs showed an intermediate disposition, although they had a stronger relationship than A+T-rich with low recombination rate regions. No significant difference was observed regarding associations between G+C composition and measures of conservation. Overall it thus seemed that G+C content was the most important factor associated with microsatellite distributions in the human genome, as associations with other genomic features were probably a by-product of their relationship with G+C content.

It was unexpected that correlation values did not differ more between PSMs and NPMs, as they characterized two exclusive datasets (Appendix, Figure 2). Of particular significance, however, was the comparison between PSMs and NP3Ms, i.e. between microsatellites that appeared or were maintained only after the human-chimpanzee-rhesus divergence about 25 Myr ago (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) and microsatellites that predated the time of the last common ancestor of human, mouse/rat and dog/cow, i.e. ~98.9 Myr ago (Bininda-Emonds et al. 2007). Due to low numbers and for statistical purposes, we did not partition NP3Ms relative to their G+C composition, but an overall small association with G+C rich regions (Table 2.2) and an analysis of microsatellite composition in the different datasets (Appendix, Figure 2) showed that A+T-rich microsatellites were depleted in NP3Ms compared to other datasets.

Overall, NP3M behavior was thus similar to that of G+C-rich microsatellites; this could explain why recombination rate appeared to covary with NP3M density and why, overall, no negative correlation was found with gene density (Table 2.2). NP3Ms also appeared to be slightly, but negatively correlated with SNP density, whereas there was no such association with PSMs. This might indicate that microsatellites lying in more labile regions degenerate, and thus disappear from genomes, at a faster rate than microsatellites that lie in more constrained regions. This is concordant with the view that some regions of mammalian genomes are more "flexible", enduring many substitutions and insertions over time, whereas other regions are more "rigid" and accumulate fewer mutations (Chiaromonte et al. 2001). Accordingly, NP3Ms showed a much stronger association with conserved elements than PSMs (Table 2.2), confirming that microsatellite conservation increases in constrained regions.

2.4.5 Genomic location influence microsatellite conservation

Miller et al. (2007) demonstrated that alignability, hence conservation, of coding exons declines more slowly than the non-coding non-repetitive portion of the human genome. Our observations were in agreement with these results; Figure 2.3B shows that conservation of microsatellites in coding exons declined more slowly than conservation of microsatellites in UTRs, which in turn declined more slowly than loci in introns and intergenic regions.

To examine further how the phylogenetic extent of microsatellite conservation may depend on their location in the genome, we looked for the distribution of human microsatellites conserved in coding exons, UTRs, introns and intergenic regions (IGRs) in each of the comparison species (Figure 2.6). The vast majority of conserved microsatellites

lied in non-exonic regions. The proportions of microsatellites found in each genomic region were fairly constant for microsatellites conserved in eutherians, with ~55-60% lying in IGRs of the human genome, ~35% in introns, and ~5-10% in exons (UTRs and cds), but varied considerably for microsatellites conserved in more distant species. The decrease in conservation was slower in exonic than non-exonic regions when phylogenetic distance increased (Figure 2.3B), a pattern similar to that of evolutionarily conserved elements (ECRs, Loots and Ovcharenko 2007). However, whereas Loots and Ovcharenko observed that >75% of ECRs shared between human and non-mammalian vertebrates were in coding regions, we found that at most 35% of conserved microsatellites were in exonic regions (human-fugu comparison). Although this figure might be underestimated due to spurious alignments with distant vertebrates (see below), there is however a well known distribution bias of microsatellites towards non-exonic regions of vertebrate genomes (Tóth et al. 2000). This distribution bias possibly arose from selective pressures acting against frameshift mutations in the reading frame of codons, therefore limiting expansion of microsatellites other than tri- and hexanucleotide repeats in these sequences (O'Dushlaine et al. 2005).

Interestingly, we found a less extensive bias in broadly conserved microsatellites according to G+C enrichment: 10.5% of A+T-rich NP3Ms compared to 39.4% of G+C-rich NP3Ms were found in exons, while values were more equivalent for MSATs (9.1% *vs.* 1.3%, Appendix, Figure 6). Overall, A+T-rich microsatellites were the most abundant class of microsatellites in human sequences, but they also disappeared more rapidly than other microsatellites, which is consistent with a genomewide study of microsatellite mutability that found that A+T-rich microsatellites were both more abundant and alterable than G+C-rich microsatellites in human and chimpanzee (Kelkar et al. 2008). More generally, our

results are in agreement with the well-documented AT mutational bias, i.e. the predominance of GC→AT *vs.* AT→GC changes (Lipatov et al. 2006).

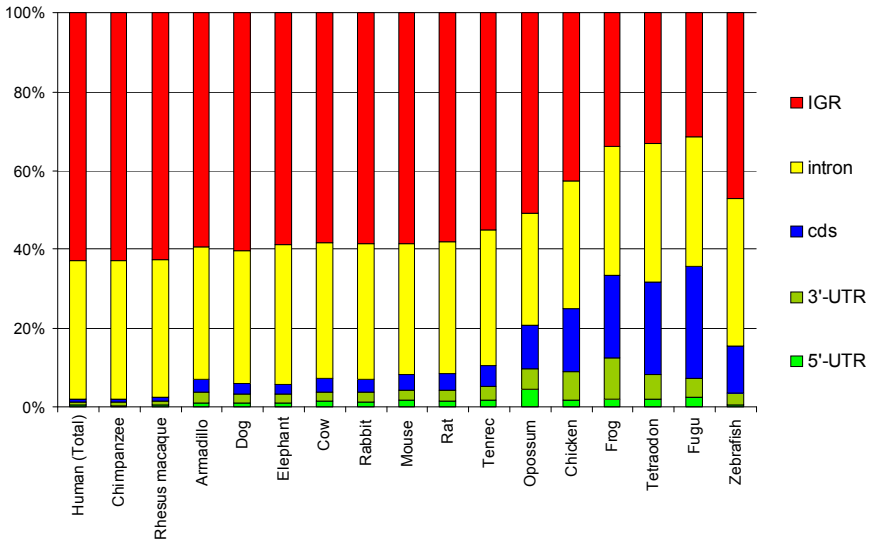


Figure 2.6: Distribution of conserved microsatellites in the human genome.

2.4.6 The reliability of large scale alignment and microsatellite data mining

Our results are only as accurate and reliable as the sequence assemblies, the genomic alignments, and the microsatellite search algorithm.

First, concordant with our experience and preliminary tests (unpublished), SciRoKo (Kofler et al. 2007) has recently been recognised as a highly performing tool to mine for perfect and imperfect microsatellites in genomic sequences (Sharma et al. 2007). As explained previously, our choice of parameters is rather conservative, so that only 'true' microsatellites are given as output.

Second, coverage and accuracy, i.e. extent of sequence gaps and errors, of the genomic assemblies available at the time and used to produce the UCSC 17-WA are variable (Table 2.1). In particular, the alignment contains 4 mammalian genome assemblies with a 2X depth coverage, namely rabbit, armadillo, elephant and tenrec, which may significantly increase the amount of false negatives in our results. According to the Lander and Waterman formula (Lander and Waterman 1988), Miller et al. (2007) calculated that a 2X assembly should include 87.5% of the bases in the genome, and a 5X assembly 99.4%. While high coverage of every genome would clearly be preferable, increasing the available branch-length with low coverage assemblies still considerably improves the accuracy of multiple genome alignments (Margulies et al. 2006; Wong et al. 2008) and of the identification of short conserved elements (Eddy 2005), and therefore improves our analysis too.

The UCSC 17-way and chain alignments are the third, and arguably the most critical (Wong et al. 2008) source of potential inaccuracies and missing data in our results. This is caused by (i) erroneous or missing genomic sequences (see above), (ii) the methodological difficulties to produce a true alignment for sequences generated from highly diverged species (Kumar and Filipski 2007), and (iii) the phylogenetic tree used to construct the 17-WA that differ slightly from the most recent understanding of evolutionary relationships between the compared species (Miller et al. 2007). Also, unlike the recently updated 28-way alignment, the generation of the 17-WA did not include filtering of pairwise alignments based on synteny (for high-quality mammalian sequences) and reciprocal best alignments (for 2X mammalian genomes). Since these advances were published only in the latest phase of our present work, we rather sought to assess the accuracy of our results *post hoc*. Fortunately, the accuracy of the 17-WA has recently been estimated through statistical inference of sequences suspiciously aligned to human

chromosome 1 (Prakash and Tompa 2007). The authors estimated that BLASTZ/MULTIZ algorithms performed well, with 9.7 % (21 Mb) of chromosome 1 identified as suspiciously assigned. Using their data, we worked out the proportion of human conserved microsatellites identified in these suspiciously aligned sequences (Figure 2.7). Results ranged from 0% (chimpanzee) to 52% (tetraodon). As expected, we observed a positive trend between the proportion of microsatellites found to be 'suspiciously conserved' in each species and sequence divergence, hence phylogenetic distance from human. There were less than 5% of human microsatellites in suspicious alignments with eutherian sequences, just over 10% with the opossum, and over 18 % with non-mammalian genomes.

Overall, we believe that our method is a robust and rapid approach for identifying human microsatellites conserved in mammals, especially in eutherians, but will suffer from badly aligned sequences when applied to more distant vertebrates. We chose to leave in our analyses all results from vertebrate comparisons because only suspicious alignments to chromosome 1 were identified to date. Overall, we recommend that these results be viewed as a preliminary attempt to characterise microsatellites in non-mammalian vertebrates and, only with particular care, be used for interpretations stemming from comparisons of incomplete 2X covered genomes.

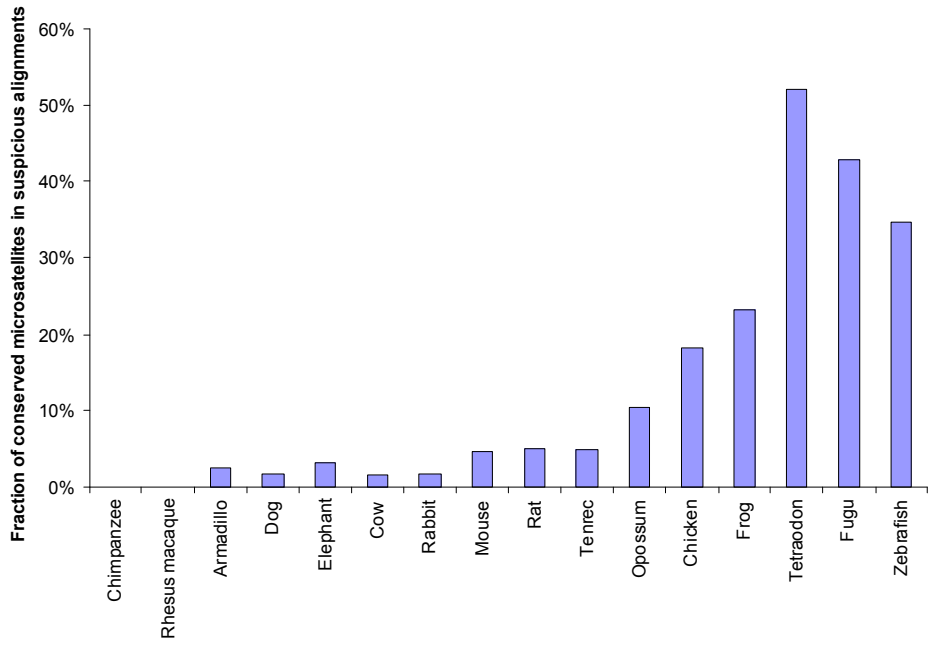


Figure 2.7: Conserved microsatellites in suspicious alignments.

2.5 Discussion

Microsatellites comprise ~3-5% of mammalian genomes (Warren et al. 2008), but very little is known about their biological significance in comparison to other genomic elements. Ironically, despite our ignorance of their role(s), if any, in the genome and an incomplete understanding of microsatellite mutational dynamics, microsatellites have been widely employed as genetic markers for almost two decades. There is therefore an obvious need for comprehensive surveys of microsatellites to explore their evolution, possible functionality, and eventually understand their place in genomes, aiding our understanding of how genomes are organised.

Surveys of the distribution of microsatellites in individual genomes (e.g. Calabrese and Durrett 2003; Subramanian et al. 2003; Warren et al. 2008) are necessary but not sufficient steps towards this understanding, as they only offer a snapshot of microsatellites

rather than providing the evolutionary time frame and evolutionary scope that genome comparisons provide to uncover mutational processes. In recent years, completion of numerous sequencing projects together with the development of new algorithms to align whole genomes have opened new perspectives to study microsatellites using a comparative approach. Here, we present the first comprehensive analysis of human microsatellite conservation in vertebrate genomes. We defined conserved microsatellites as microsatellite sequences irrespective of motif and structure that were identified at orthologous positions in different genomes. Drawing on the UCSC alignment of the human genome against the genomes of 11 mammals and 5 non-mammalian vertebrate species, we were able to find all human microsatellites that were conserved above the species, genus, group or even family level. Our findings therefore significantly extend the scope of previous reports of microsatellite conservation and the sporadic identification of microsatellites conserved above the genus level, in mammals (e.g. Schlötterer et al. 1991; Moore et al. 1998) and other vertebrate species (e.g. FitzSimmons et al. 1995; Rico et al. 1996).

Our initial microsatellite scans in the 17-WA sequences confirmed that microsatellite content varies greatly among genomes (Warren et al. 2008), and that these differences are species-specific rather than correlated with phylogenetic distance. We have previously suggested that the evolution of microsatellites in genomes could be described by birth, growth, degeneration, death and re-birth cycles (Buschiazzo and Gemmell 2006). If rates of birth and death events were high enough to promote a large turnover of microsatellite sequences in vertebrate genomes, then it would be indeed unexpected to find a significant number of conserved microsatellites, even between closely related species. On the other hand, one could expect high conservation of microsatellites if microsatellite degeneration and loss were rare events. As a consequence of the complexity and heterogeneity of microsatellite mutational dynamics, there is to date no theoretical

development to estimate the life expectancy, thus the turnover, of microsatellites above the species level (Stephan and Kim 1998). Our results provide new information on the mode and tempo of conservation of microsatellites and therefore could be fundamental for such developments.

We found that of 696,016 microsatellites identified in aligned human sequences, 85.39% were conserved in at least one species, 28.65% in at least one non-primate species and 5.98% in at least 3 non-primate species. On the whole, this decline of conservation appeared exponential as a function of evolutionary distance, and did not necessarily always depend on time of divergence alone. For example, 7.68% of human microsatellites were conserved in the elephant genome (~100 Myr ago), but only 7.17% in the mouse genome (~90 Myr ago). While the mouse genome is known for its lability (Waterston et al. 2002; Lindblad-Toh et al. 2005), little is known about the elephant genome despite a surprising similarity to the human genome, evident from our results of microsatellite conservation but also from its high representation in the 17-WA where it aligns to 37.39% of the human genome. The exponential decline of microsatellite conservation is consistent with sequence loss through random genetic drift, and thus supports the general view that most microsatellites evolve neutrally and would therefore be maintained only by chance. We also found that the chromosome and megabase distribution of microsatellites were very similar regardless of the extent of conservation, providing further support for the neutral degradation of microsatellites. However, broadly conserved microsatellites (NP3Ms) had a slightly different distribution, with specific megabase portions of the human genome containing more of these microsatellites than expected, suggesting that at least some microsatellites were under selective constraints. Similarly, a decline of G+C-rich microsatellites in exons (cds and UTRs) was found to occur more slowly than that of A+T-rich microsatellites in intron and IGRs.

These findings raise questions as to why microsatellites are conserved in distant species, and why microsatellites in different genomic locations are maintained to different extents. Coding microsatellites may be subject to purifying selection as they might be important for protein structure and protein-protein interactions (Hancock and Simon 2005) or to indirect selection as a source of adaptive evolution (Wren et al. 2000; Fondon and Garner 2004), and some microsatellites have been shown to be selected for their folding potential rather than their primary sequence in 5'-UTRs (Riley et al. 2007). Although it is not clear what the function of most non-exonic microsatellites is, there is clear evidence that at least some are acting as regulators of gene expression (Kashi and King 2006), suggesting that non-coding microsatellites could also be indirectly selected for mutability. Indeed, the genetic variation provided by microsatellites may be advantageous and may vary (and evolve) independently from otherwise low average nucleotide-substitution rates (Kashi and King 2006).

Conserved microsatellites therefore provide exciting possibilities to help single out those loci that may be actively selected for functionality, but there might be a need for further data and theoretical developments (i.e. statistical tests) to reliably distinguish between mere retention (neutral) and active conservation (selection).

Conserved microsatellites also allow the exploration of the mutation dynamics of microsatellites above the species level, an approach that has been rarely used to date (Zhu et al. 2000; Kelkar et al. 2008). In particular, further investigation is needed to tease out structural changes among orthologous microsatellites (Chapter 5), e.g. whether there are motif changes (Riley et al. 2007), whether compound structures arise from simple structures, and whether there are interspecies and intraspecies variations in length and/or mutability (Laidlaw et al. 2007; Kelkar et al. 2008). Mutable conserved microsatellites will also prove particularly useful to develop and implement transferable PCR primers. Indeed,

one disadvantage of microsatellites as genetic markers is that cross-species studies needs substantial preparation (Barbara et al. 2007); provided that priming sites are also conserved between species of interest, conserved microsatellites overcome this limitation and are therefore a valuable resource for cross-species applications in population genetics, comparative molecular ecology and gene mapping.

2.6 Acknowledgments

Help from V. Mencl was crucial to optimize our computational work at the University of Canterbury. A. Bagshaw, D. King, R. Kofler, R. Sainudiin, J. Tylianakis, and I. Vargas-Jentzsch provided suggestions to improve the manuscript. Genomic intervals of suspicious alignments were kindly provided by H. Prakash and M. Tompa.

Chapter 3

3 Evolutionary and phylogenetic significance of microsatellites conserved in platypus and other vertebrates

3.1 Abstract

Microsatellite sequences are generally considered to exhibit complex mutation dynamics and to be highly labile in an evolutionary sense, with most loci conserved only among closely related species. The identification of all human microsatellites conserved in at lest one of 16 vertebrate species (Chapter 2) showed that most microsatellites are indeed lost rapidly, but that a large subset is retained over a large evolutionary scale. Drawing on the recent publication of the first monotreme genome, that of platypus, and the production of an associated multiple genome alignment (6-way alignment, i.e. 6-WA) containing three mammals (opossum, human, mouse) and two non-mammalian vertebrates (chicken, lizard), it was timely to inspect wide-ranging microsatellite conservation in the monotreme lineage. Most platypus microsatellites were conserved in one species, as expected by the large evolutionary distance that separates platypus with other species, with platypus sharing more microsatellites with mammals than with reptiles. Within mammals, unexpectedly, more platypus microsatellites had orthologues in the opossum genome than in that of either human or mouse, which was at odds with the current view that monotremes diverged from a lineage containing both eutherians and marsupials (Theria hypothesis). The phylogenetic significance of microsatellite conservation was further investigated through Bayesian and maximum parsimony tree reconstruction, using presence/absence data of microsatellites conserved in 18 species, including platypus. Although models of evolution implemented in current phylogenetic reconstruction algorithms are not tailor-made for microsatellite data, we were able to construct reasonably good vertebrate phylogenies, with two of our three reconstructions supporting the Theria hypothesis. Identifying the fraction of platypus microsatellites conserved in the 6-WA not only helped understand better the evolutionary dynamics of microsatellites in vertebrates,

but also provided ground for theoretical development in phylogeny-based analyses of microsatellite data.

3.2 Introduction

The platypus, an animal native to Australia, is one of only three extant species of monotremes, together with two echidna species, that form the mammalian subclass Prototheria. Monotremes are arguably the most unusual of all extant mammals, owing to their unique combination of ancestral reptilian features, such as egg-laying, and derived mammalian features, including lactation. Recently, the platypus genome was added to the pool of published sequenced genomes (Warren et al. 2008), and comparisons of the platypus genome to the genomes of other mammals and of the chicken revealed new insights into early mammalian evolution. Although the platypus genome exhibits specific monotreme features, such as unique expansions of certain protein and RNA families, and an unusually high G+C fraction (45.5% *vs.* 40.7% for human and similar values for other mammals and chicken), there is strong evidence that platypus chromosomes mirror the mixture of ancestral reptilian and derived mammalian biological features. For example, platypus X chromosome sequences share no homology with eutherian X sequences, but show substantial homology to the chicken Z chromosome, suggesting that the platypus sex chromosome system evolved directly from a bird-like ancestral system, while the therian (marsupials and eutherian mammals) sex chromosome system evolved independently after divergence from a common ancestor. In addition, the mean microsatellite coverage of platypus genomic sequences into chromosomes (2.67±0.34%) is significantly lower than observed for all mammalian genomes sequenced so far, and most similar to that observed in chicken. Furthermore, platypus sequences contain a higher proportion of A+T-rich

microsatellites compared to other vertebrate genomes, but the abundance distribution has more in common with chicken than with mammals. On the other hand, about half of the platypus genome can be annotated as interspersed repeats, a feature that is similar to available mammalian genomes, but in contrast to that of chicken.

All together, these characteristics suggest that the platypus genome, mirroring the animal's morphology, is an amalgam of reptilian and mammalian features. As such, the platypus is an important evolutionary link between reptiles and other mammals. Phylogenetic analyses based on nuclear genes and indels showed that monotremes diverged early from the metatherian-eutherian lineage (Theria hypothesis, van Rheede et al. 2006; Warren et al. 2008), as opposed to earlier mitochondrial DNA sequence comparisons that placed marsupials and monotremes into a same clade (Marsupionta hypothesis, Janke et al. 1996; Janke et al. 2002). The most recent analysis based on fossil and molecular data estimated the monotreme-theria divergence at ~166 Myr ago (Bininda-Emonds et al. 2007; Warren et al. 2008).

Despite this key position at the base of the mammalian phylogeny, monotremes have remained missing from comparative studies of the human genome. However, the inclusion of a monotreme in comparative analyses adds significant value by bridging a divide between comparison with chicken, which is too distantly related to human for useful comparisons to be made, and other eutherians, which are too closely related to human to provide great evolutionary insight. The UCSC whole-genome 17-WA, with an otherwise remarkable array of 11 eutherian and one marsupial genomes, illustrates this shortfall: with the sequence of the platypus missing, misalignment or absence of other otherwise alignable sequences between chicken and mammalian genomes may have occurred.

In Chapter 2, I presented the first comprehensive study of microsatellite conservation in vertebrate genomes using the then only available 17-WA. With accessibility to the platypus genomic sequence and to an associated 6-WA alignment restricted to the genomes of human, mouse, opossum, chicken and lizard, it was timely to inspect and describe the retention of vertebrate microsatellites in the monotreme branch, and integrate these results into the 17-WA analysis framework. In addition to providing an extended framework for future work, this dataset offered a unique opportunity to reconstruct a vertebrate phylogeny based on binary data, indicative of the presence/absence of microsatellites in compared genomes. Although resulting topologies were reasonably accurate and demonstrated the potential of the approach, contradictions with the currently assumed mammalian tree highlighted the need for improved theoretical models of microsatellite mutational dynamics and evolution above the species level.

3.3 Materials and Methods

3.3.1 Vertebrate sequences

The 6-WA available on the UCSC Genome Browser for the platypus genome was downloaded by anonymous FTP from ftp://hgdownload.cse.ucsc.edu/goldenPath/ornAna1/multiz6way. MAF-formatted blocks were extracted and converted to FASTA format using a stand-alone version of Galaxy (Giardine et al. 2005) downloaded from http://main.g2.bx.psu.edu/. Due to its large size, the alignment file was split in half. Gaps were removed using the degapseq module from the EMBOSS 5.0 package (Rice et al. 2000). Table 3.1 shows the characteristics of each assembly.

Table 3.1: Species in the 6-WA. Cover: genome sequence coverage; Dist 28-WA: branch lengths retrieved from 28-way UCSC data; Diverg. time: divergence time from platypus lineage (Myr ago), from (Bininda-Emonds et al. 2007); Platypus align: aligned fraction of the platypus genome.

Scientific name	Common name	UCSC ID	Cover	Dist 28-WA	Diverg. time	Seq size (Gb)	In 6-WA (including overlaps)	Platypus align
Ornithorynchus anatinus	Platypus	ornAna1	6x	0	0	1.84	1.12	100%
Monodelphis domestica	Opossum	monDom4	6.5x	0.9040	166.2	3.50	0.32	29.02%
Homo sapiens	Human	hg18	Fin	0.9738	166.2	2.86	0.24	21.90%
Mus musculus	Mouse	mm8	Fin	1.1724	166.2	2.60	0.15	13.80%
Gallus gallus	Chicken	galGal2	6.6x	1.0020	326.0	1.05	0.12	10.87%
Anoles carolensis	Lizard	anoCar1	6.8x	1.1209	326.0	1.74	0.09	8.31%

3.3.2 Microsatellite search and classification

Perfect and imperfect microsatellites (motif length: 1 to 6 bp) were searched in ungapped sequences using SciRoKo 3.1 (Kofler et al. 2007) with fixed penalty parameters (score: 12, mismatch penalty: 4, SSR seed min. length: 3, SSR seed min. repeats 3, max. mismatches at once: 3). Genomic intervals of microsatellites in each vertebrate genome were recorded with block number, standardized repeat motif, array length, and number of imperfections. Overlapping intervals (5 bp minimum cut-off) of non-platypus species were removed. Intervals overlapping with repeats other than simple repeats or low complexity sequence (Smit et al. 1996-2007) were also discarded. Repeat data were retrieved from the UCSC Table Browser (Karolchik et al. 2003). Microsatellites were classified as 'simple', 'compound', 'linked' or 'mixed', as described in Chapter 2. This series of operations produced, in platypus, a dataset of 351,562 microsatellites, representing a total length of 8.02 Mb (0.72%% of total platypus sequences in the 6-WA).

3.3.3 Microsatellite conservation

Genomic intervals of non-platypus microsatellites were converted to the ornAna1 platypus assembly using the liftOver utility and chain files (Kent et al. 2003) available at the UCSC Genome Browser (Karolchik et al. 2003). The fraction of platypus microsatellite positions that overlapped with any of the converted microsatellite positions indicated conserved sites.

3.3.4 Integration into the 17-WA framework

Platypus microsatellites identified in the 6-WA sequences were converted to the human UCSC hg18 assembly using liftOver, and overlaps with microsatellites identified as conserved in the 17-WA sequences were recorded. Microsatellites either lost or conserved in human were considered.

3.3.5 Phylogenetic inference

The general approach was inspired by an earlier work on phylogenetic reconstruction based on gene content (Huson and Steel 2004). A binary matrix was generated for conserved and non-conserved microsatellites in the 17-WA of human autosomes, including platypus microsatellites: conserved (present) and non-conserved (absent) microsatellites were coded 1 and 0, respectively, at each locus and for each species. An additional dataset consisting of 0 only was included to represent a virtual outgroup species. This resulted in a 19-species matrix containing 833,147 concatenated characters.

Bayesian analyses were conducted with parallel MrBayes (vers. 3.1.2, Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Altekar et al. 2004) and parallel BayesPhylogenies (Pagel and Meade 2004) using the BlueFern supercomputing resources at the University of Canterbury (New Zealand) in the BeSTGrid environment. The parallel processes ran 10 million generations of Markov-Chain Monte Carlo on four chains. For BayesPhylogenies, variable rates of gain and loss were assumed (morphological model m2p) and the virtual outgroup species was set as outgroup. Empirical character frequencies were: 0= 85%, 1= 15%. From the resulting set of 1001 trees, a consensus tree (consensus level = 50) was obtained with averaged branch lengths and posterior probabilities of clades using Treefinder (Jobb 2007). For MrBayes, the restriction site (binary) model with equal rates of change was employed and a coding bias ('noabsencesites') was assumed as characters that are absent (0) in all species cannot be observed.

A maximum parsimony heuristic search was also conducted in PAUP* vers. 4.0b10 (Swofford 2002) on a Pentium 4 3.2 GHz (2.87 GB of RAM), consisting of 10 random addition sequence replicates using the tree-bisection-reconnection (TBR) branch-swapping algorithm. In addition to searching for the optimal tree, a bootstrap analysis was conducted based on 2,000 pseudoreplicates with 10 random addition sequence replicates per pseudoreplicate. The TBR algorithm was also used in this analysis.

Base trees were drawn using FigTree v1.1 (Rambaut 2006-2008).

3.4 Results

3.4.1 Microsatellites in the alignment

The common ancestor to all mammals diverged from reptiles ~320 Myr ago (Blair and Hedges 2005), and monotremes diverged from the therian lineage leading to marsupials and eutherians ~170 Myr ago (Bininda-Emonds et al. 2007; Warren et al. 2008). Platypus features are therefore expected to be more similar to those of other mammals than of reptiles, a distinction that is reflected in the alignment of its genome. Indeed, we found that the aligned fraction of the platypus genome (alignability) was higher for mammalian genomes (14%-29%) than for reptilian genomes (8%-11%, Table 3.1). Consistent with the homogeneous distribution of microsatellites in genomes, the numbers of microsatellites identified in the different genomes followed a trend similar to that of alignability (Figure 3.1A). We found an overall positive relationship between evolutionary distance from platypus, alignability and the number of microsatellites identified in aligned sequences (Table 3.1, Figure 3.1), reminiscing results presented in Chapter 2 for the alignment of the human genome against 16 vertebrate genomes. Exhibiting the highest substitution rate among all species (Miller et al. 2007), mouse sequences were less represented in the alignment and contained less microsatellites compared to human and opossum sequences, satisfying the expectation, but alignability and microsatellite content was higher is mouse than in reptiles, suggesting that evolutionary rate of genomes and genetic distance does not explain all the observed differences. Rather, the structures of reptilian and mammalian genomes have diverged significantly since the split at the time of the most common ancestor, and even though the platypus genome retained reptilian features, a larger fraction

aligns to genomes of other mammals, and probably more platypus microsatellites are shared with other mammals, including mouse, than between platypus and reptiles.

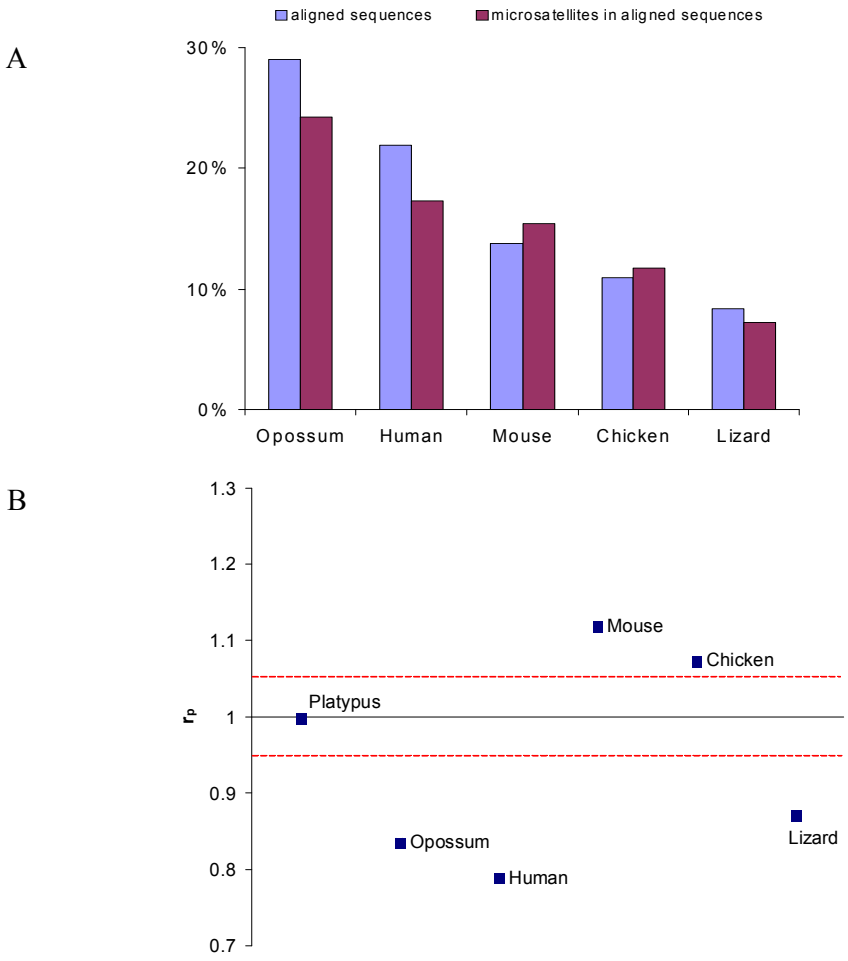


Figure 3.1: Species-specific microsatellite enrichment. (A) Alignability and microsatellite abundance in vertebrate genomes relative to the platypus genome. (B) Scatter plot showing the ratio (r_p) of percentage of alignment to percentage of microsatellite conservation relative to platypus. Dotted lines represent a 5% significance threshold.

The ratio of the percentage of microsatellite abundance to alignability was calculated for each species (Figure 3.1B) to assess microsatellite enrichment compared to platypus. Values for therian species confirmed the results presented in Chapter 2; human sequences were impoverished in microsatellites compared to opossum and especially mouse sequences. Comparatively, platypus sequences showed an intermediary enrichment

between opossum and mouse sequences; this result apparently contradicts genome-scale microsatellite scans (Warren et al. 2008) that showed that the platypus genome contained fewer, albeit long, microsatellites than any other mammalian genomes analysed.

Using the platypus genome as the reference genome in the alignment might have created an ascertainment bias that resulted in the identification of relatively more microsatellites in platypus compared to sequences from the other species. A further matter of contention was found, as chicken sequences had the highest microsatellite enrichment behind mouse sequences in the 6-WA, whereas previous studies reported the lowest microsatellite content in chicken sequences among analysed species (Warren et al. 2008, Chapter 2). Although the reason for this discrepancy is unclear, the nature of those chicken sequences that aligned to the platypus genome might be different from the overall genomic content; particularly, they seem to contain proportionally more microsatellite-rich regions.

3.4.2 Interchromosomal distribution of conserved platypus microsatellites

The interchromosomal distribution of microsatellite conservation may be affected by the karyotype, assembly quality and alignability of the platypus genome. The platypus karyotype consists of 26 chromosome pairs in both sexes, including five sex chromosome pairs, with a size distribution similar to the bimodality observed in reptilian macro- and microchromosomes. Unfortunately, the platypus genome assembly (UCSC ornAna1) suffers from incomplete sequence coverage (1.84 Gb of completed sequence *vs*. an estimated size of 2.4 Gb) and significant fragmentation (0.79 Gb of contigs and ultracontigs are not assigned to any chromosome). The combination of higher G+C content and the increased density of interspersed repeats compared to previously characterized

genomes are thought to be at the origin of the difficulties to sequence and assemble the entire platypus genome (Warren et al. 2008). There was a striking size distribution bias among sequences represented in the 6-WA (Figure 3.2): sequences assembled in chromosomes 6, 7, X1, X2 and X3 were aligned in their near totality, whereas sequences from other chromosomes were only partially aligned (21-28% of the ungapped length). Discussing the foundation of this bias is beyond the scope of the present analysis.

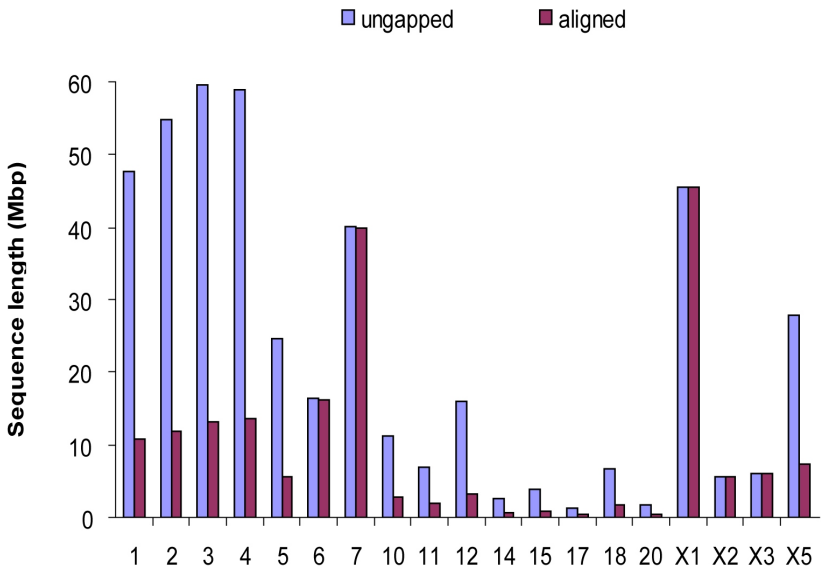


Figure 3.2: Sequence length (bp) per chromosome.

A direct consequence of this bimodal alignability is illustrated in Figure 3.3; well-represented chromosomes showed a much lower proportion of conserved microsatellites as a function of their total length compared to partially represented chromosomes. Although this suggests that the proportion of conserved microsatellites decreases with the length of sequence aligned, no significant relationship was found between alignability and the proportion of microsatellite conservation when highly aligned chromosomes were excluded (Spearman's rank correlation test, $P>0.05$), which suggests that microsatellite conservation is fairly heterogeneous among chromosomes. This finding differs from conclusions drawn from the analysis of microsatellite conservation in the human genome

(Chapter 2). Although it could be suggested that there is indeed a distribution bias in the amount of microsatellite conservation among platypus chromosomes, this hypothesis should only be tested once the platypus genome assembly reaches a complete or almost complete stage.

3.4.3 Phylogenetic extent of conservation

The species used in our comparison examining the phylogenetic conservation of microsatellite loci are distantly related to each other, thus our expectation was that platypus microsatellites conserved in several species should be relatively rare, whereas platypus microsatellites shared with a single species should exist in proportionally larger numbers. Of 352,034 microsatellites identified in platypus sequences, 20,441 microsatellites, i.e. 5.81 %, were conserved in at least one species, including 75% conserved in exactly one species and 10% only conserved in three or more species (Figure 3.3). Specifically, the fraction of platypus microsatellites identified in the 6-WA sequences that are conserved in other species was 0.77% in lizard, 1.19% in chicken, 1.81% in mouse, 1.85% in human and 2.55% in opossum. This decrease in the proportion of conserved microsatellites as the number of species increases supports the neutrality of most microsatellites and the loss through random genetic drift over the large evolutionary distance separating platypus from other species in the comparison (Chapter 2).

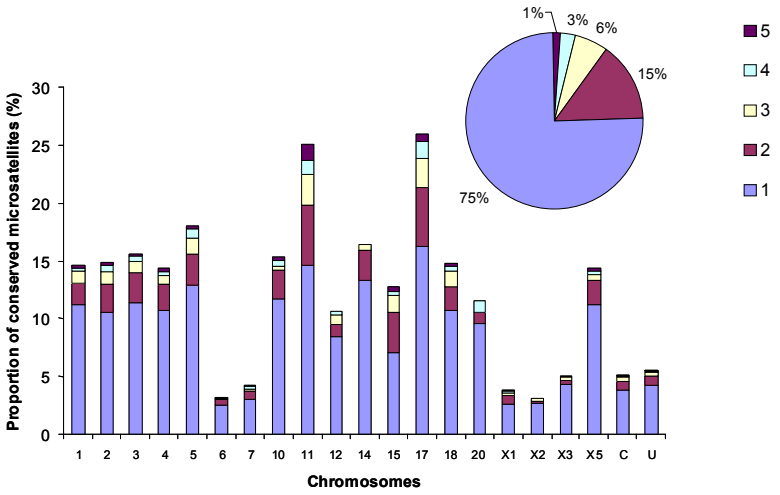


Figure 3.3: Chromosome distribution of conserved microsatellites. Color coding indicates the proportion of initial microsatellites that were found conserved in groups of species of increasing size, relative to the total number of microsatellites in each chromosome (*bar chart*) and to the genomewide number of conserved microsatellites (*pie chart*). C: contigs, U: ultracontigs.

As expected from phylogenetic relationships among amniotes and the position of platypus within the mammalian clade (Warren et al. 2008), more platypus microsatellite loci were conserved in mammals than chicken and lizard (Figure 3.4). However, curiously, more platypus microsatellites were conserved in opossum than the other mammals. First, the species-specific distribution of platypus microsatellites conserved in exactly one, two, three, four and all five species (Figure 3.4A) indicates that species combinations including lizard and/or chicken comprised fewer microsatellites than combinations including human, mouse and especially opossum. Second, drawing on these count data, a conservation profile graph was produced (Figure 3.4B). As explained in Chapter 2, a profile skewed to the left indicates a species that share microsatellites relatively exclusively with platypus (e.g. opossum), whereas a profile skewed to the right indicates a species that shares microsatellites that are broadly conserved (e.g. chicken and lizard). Mouse and human showed intermediary profiles.

Altogether, the analysis of the extent of phylogenetic conservation of platypus microsatellites in vertebrates therefore suggested a closer relationship between monotremes and marsupials. It is however unclear how useful evolutionary interpretations based on extent of microsatellite conservation might be versus other data supporting the Theria hypothesis (van Rheede et al. 2006; Warren et al. 2008).

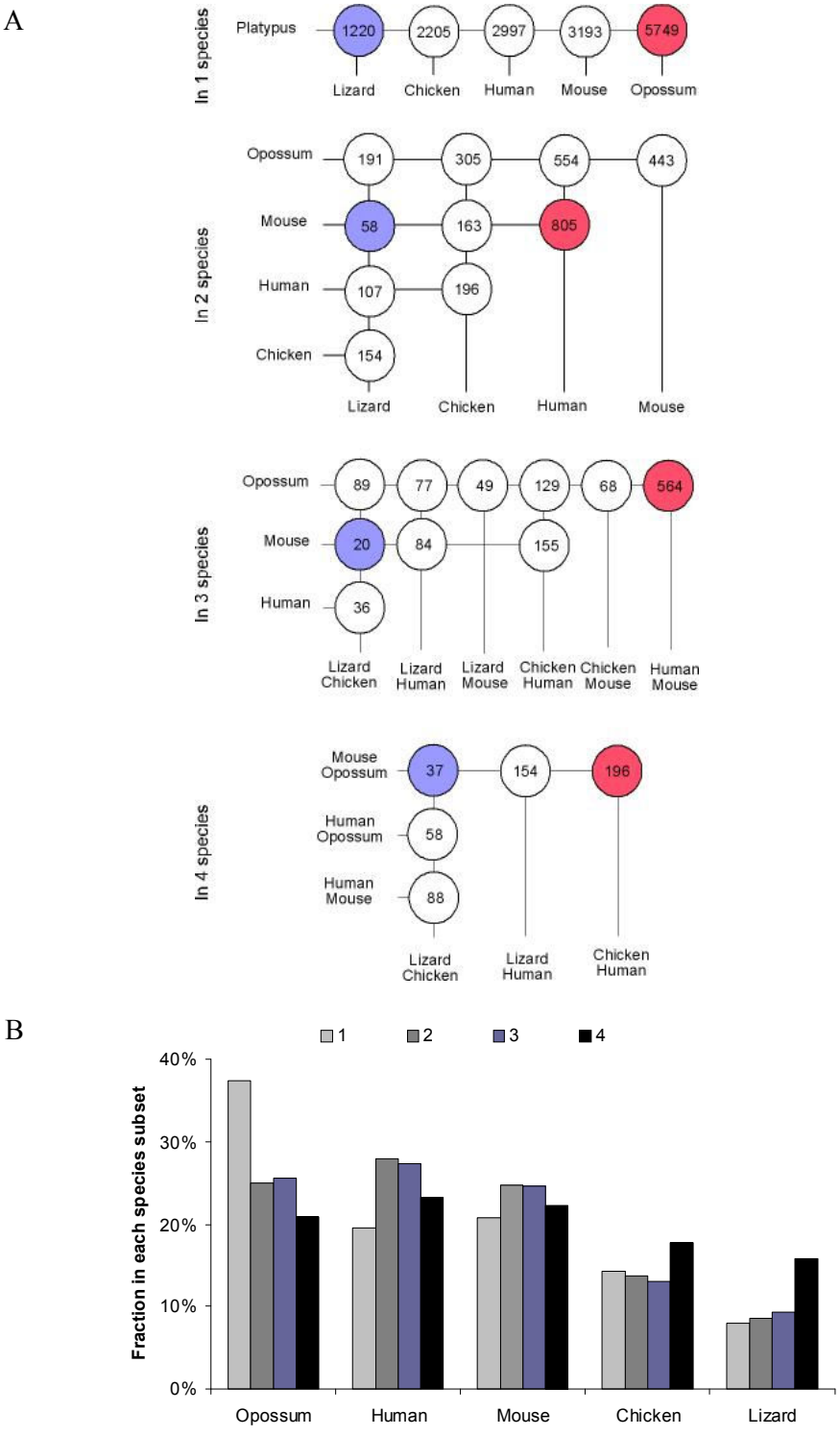


Figure 3.4: Extent of platypus microsatellites conservation. (A) Microsatellite counts in every possible combination of species. Minimum and maximum are shown in blue and red circles, respectively. (B) Conservation profile of platypus microsatellites in each comparison species. The range of conservation represents the range of species that share the same platypus microsatellites, from exclusive (1 species) to wide (5 species). Bar plots of identical range add up to 100%.

3.4.4 Phylogenetic reconstruction

To examine further the phylogenetic significance of conserved microsatellites in vertebrates, Bayesian and maximum parsimony (MP) phylogenetic tree reconstructions were used with the extensive binary data provided by the presence/absence of microsatellites conserved in vertebrate genomes, a strategy that has already been implemented on gene data using maximum likelihood and especially Dollo parsimony (Huson and Steel 2004).

Platypus microsatellites were incorporated into the framework created by the 17-WA analysis to broaden the range of vertebrate species to 18 taxa, and a virtual outgroup species was adjoined (see Methods).

When all microsatellites conserved in at least two of the 18 species were counted and encoded, the result was a matrix containing 833,147 concatenated characters. The mean composition of state characters in this matrix was: 0 = 85% and 1 = 15%. Table 3.2 details the state character composition for each taxon. A strong bias of microsatellite presence towards human and species closely related to human was expected because the human genome was used as reference in the 17-WA. Consequently, all human microsatellites conserved with any of the compared species were identified, but not all microsatellites conserved among other species were identified because sequences that did not align with the human genome were self-excluded from the analysis. The inclusion of all microsatellites identified in aligned non-human sequences but not retained in the human genome only minimized this bias (Table 3.2).

The optimal phylogenetic branching patterns among vertebrate taxa using two Bayesian and one MP analyses of the microsatellite data are shown in Figures 3.5-7. The underlying assumptions (probability of forward change, i.e. birth, much lower than reverse

change, i.e. death) and strenuous computing requirements (randomization of the order of the 19 species) of the Dollo parsimony method were incompatible with our extensive microsatellite data, which is why the method was not used.

First, Figure 3.5A depicts the optimal Bayesian tree assuming variable rates of gain and loss (m2p model, Pagel et al. 2004); the dramatic elongation of primate branches relative to other branches, caused by the character state distribution bias, is clearly illustrated. The tree was rooted on the virtual outgroup species, explaining the gradual increase of branch lengths from the root of the tree to the primate clade. Figure 3.5B shows the topology of the same tree compared to that of a tree derived from the analysis of the 28-WA (Miller et al. 2007). This tree represents the current authoritative view on the contentious phylogenetic relationships among major placental groups: a clade (Atlantogenata) composed of Xenarthra (armadillo) and Afrotheria (elephant and tenrec) is a sister group of all other crown placental mammals (Boreoeutheria), among which Rodentia (mouse and rat) groups with Lagomorpha (rabbit) to form a clade (Glires) closer to Primates (human, chimpanzee and rhesus macaque) than Laurasiatheria (dog and cow). In addition, the authoritative tree favours the Theria hypothesis, placing Monotremata at the base of the mammalian phylogeny.

Table 3.2: Presence/absence state distribution of microsatellites in 18 vertebrate genomes

		Presence		
Species	Absence	In all	In human	Not in human
Human	263,484	569,663	-	-
Chimpanzee	298,296	534,851	501,783	33,068
Rhesus	450,477	382,670	318,073	64,597
Mouse	702,603	130,544	41,014	89,530
Rat	714,525	118,622	35,703	82,919
Rabbit	765,673	67,474	38,595	28,879
Dog	673,539	159,608	71,136	88,472
Cow	705,756	127,391	55,238	72,153
Armadillo	757,971	75,176	31,762	43,414
Elephant	721,621	111,526	43,801	67,725
Tenrec	778,410	54,737	21,251	33,486
Opossum	808,816	24,331	9,849	14,482
Platypus	822,832	10,315	5,978	3,341
Chicken	827,723	5,424	1,907	3,517
Frog	829,697	3,450	1,288	2,162
Tetraodon	829,015	4,132	1,278	2,854
Fugu	830,069	3,078	998	2,080
Zebrafish	829,359	3,788	1,736	2,052
Outgroup	833,147	0	0	0

Although bearing strong support for all nodes, the m2p microsatellite-based topology showed significant differences compared to the consensus topology: (i) monophyly for Atlantogenata was not supported; (ii) Rodentia was placed at the base of the Boreoeutheria, and Glires were therefore paraphyletic; (iii) the Marsupionta hypothesis was supported; (iv) outside Mammalia, birds (chicken) and amphibians (frog) were monophyletic.

Figure 3.6 shows the optimal Bayesian topology assuming equal rate of change in a restriction site model. The topology was more similar to that of the current authoritative phylogeny compared to the m2p Bayesian analysis, although (i) rodents were placed at the base of the Placentalia, but monophyletic with Lagomorpha; (ii) dog and cow were split, with dog closer to Primates; (iii) the Atlantogenata monophyly was not supported; (iv) six nodes in the placental subtree were weakly supported (Bayesian posterior probabilities < 0.85). Importantly, however, this Bayesian analysis supported the Theria hypothesis, placing platypus at the base of the mammalian phylogeny.

Figure 3.7 depicts the optimal branching pattern using a MP heuristic search and bootstrap analysis. The definition of the mammalian phylogeny was greatly improved compared to both Bayesian analyses; with the exception of rodents placed at the base of the Euarchontoglires (i.e. Glires were paraphyletic), the topology exactly matched that of the current authoritative phylogeny of these 18 vertebrate taxa. All but three nodes had MP bootstrap equal to 100%; one of the exceptions concerned the platypus/opossum branching, with 62% of trees found in the bootstrap analysis supporting the Theria hypothesis.

Overall, tree reconstruction of the vertebrate phylogeny based on the presence/absence of microsatellites in 18 genomes was more successful with an MP approach than using two Bayesian implementations, i.e. MP trees were in better agreement with the currently authoritative phylogeny. The place of mouse and rat appeared particularly unstable and consistently in contention with the consensus view, suggesting a complex mutation history of microsatellites in the rodent lineage that is reminiscent of major rodent-specific rearrangements at the genome scale (Waterston et al. 2002; Gibbs et al. 2004; Ferguson-Smith and Trifonov 2007). The best two of the three analyses, i.e. Bayesian analysis using m2p model and MP, were in agreement with the Theria hypothesis and the current view on the place of platypus at the base of the mammalian phylogeny, suggesting that the integration of new microsatellite conservation data into the 17-WA framework is a successful strategy.

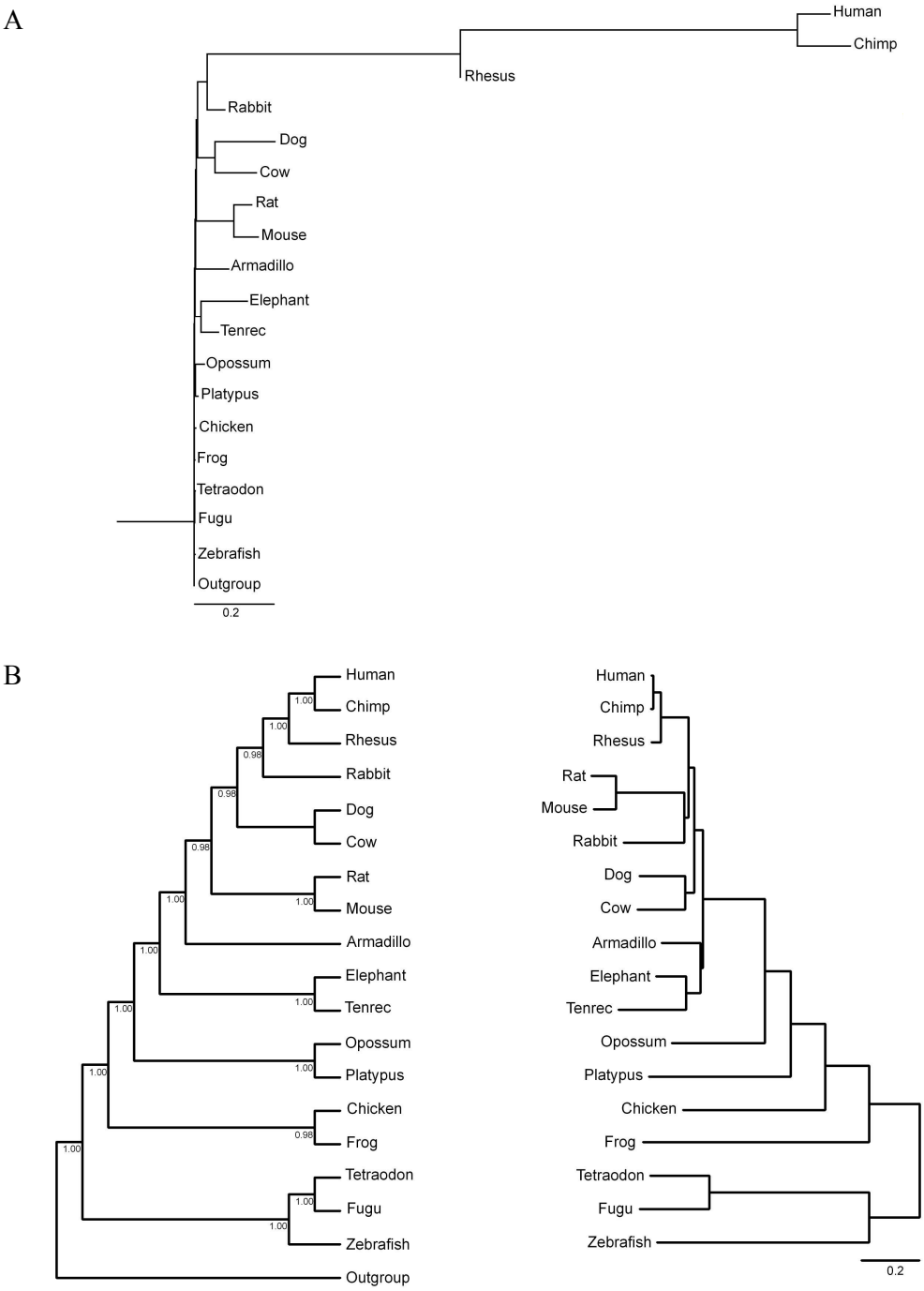


Figure 3.5: Vertebrate phylogeny inferred from Bayesian analysis using a morphological model with variable rates of change. (A) Optimal tree topology and branch lengths obtained by Bayesian analysis of presence/absence microsatellite data in 18 vertebrate species genome and one virtual outgroup species. (B) *Left:* optimal tree topology obtained by Bayesian analysis of presence/absence microsatellite data. Bayesian posterior probabilities equal 1.00 for all nodes with the exception of four nodes (=0.98). *Right:* Authoritative tree topology and branch lengths from an independent analysis in 28 vertebrates based on substitutions at 4D sites (Miller et al. 2007). Only species of comparative interest are displayed.

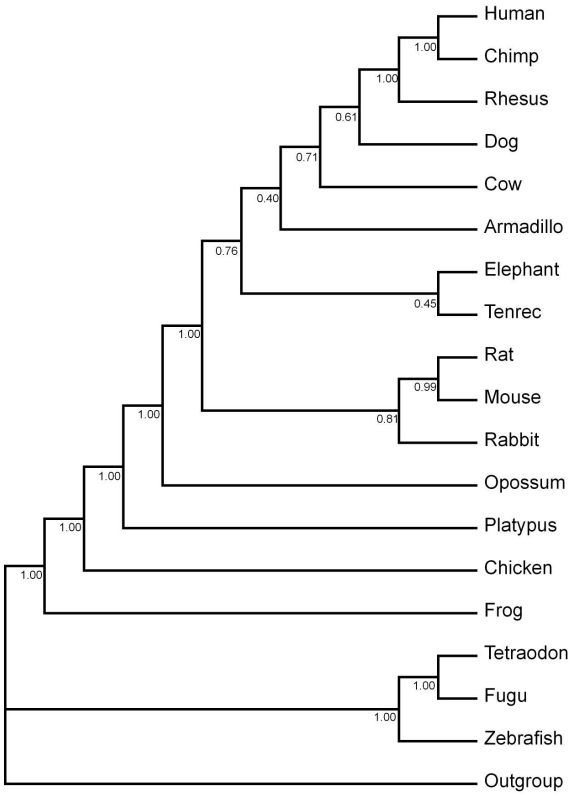


Figure 3.6: Vertebrate phylogeny inferred from Bayesian analysis using a restriction site (binary) model with equal rate of change.

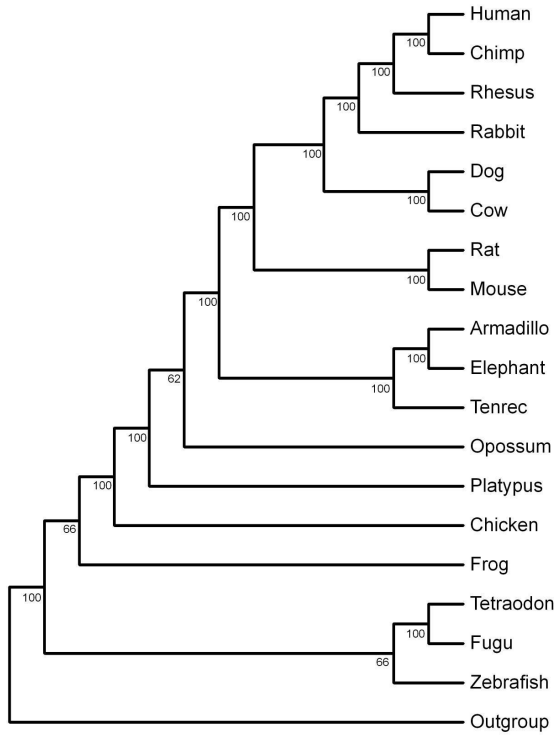


Figure 3.7: Vertebrate phylogeny inferred from MP analysis of microsatellite binary data using a heuristic search and the TBR branch-swapping algorithm.

3.5 Discussion

The first monotreme genome, that of platypus, has recently been added to the pool of genomes at the disposal of the research community (Warren et al. 2008), providing an advanced tool that was missing from past comparative analyses of the human and other vertebrate genomes. Echoing the amalgam of ancient reptilian and derived mammalian biological features that make monotremes such extraordinary animals, the platypus genome appears to share characteristics with bird and other mammalian genomes. Microsatellite coverage, for example, is significantly lower in the platypus genome than all other mammalian genomes sequenced to date and most similar to that observed in chicken. In addition, the A+T distribution bias of platypus microsatellites also observed in mammalian and reptile genomes resembles more that of reptiles (Warren et al. 2008).

The identification of the fraction of platypus microsatellites conserved in related vertebrate species adds another line of evidence corroborating this mixture of mammalian and reptilian features in platypus: of 20,441 microsatellites conserved in at least one species, 17.5% are found in reptiles only, 70.0% in mammals only, and 12.5% were found in both groups. As expected, these results suggest a closer relationship with other mammals than with reptiles.

Among mammals, significantly more platypus microsatellites are shared with opossum than human or mouse, suggesting a closer relationship of monotremes with marsupials (Marsupionta hypothesis). These patterns of microsatellite conservation seem in disagreement with the current view (Theria hypothesis) that monotremes diverged from the therian lineage (van Rheede et al. 2006; Bininda-Emonds et al. 2007) leading to placentals

(e.g. human and mouse) and marsupials (e.g. opossum). However, raw microsatellite counts may not represent well the evolutionary history of microsatellites on the mammalian phylogeny.

The method of choice to address this issue would naturally involve phylogenetic reconstruction, but this raises methodological difficulties for microsatellite-based approaches. Several methods using microsatellite-based genetic distances have been developed to reconstruct population/species phylogenies (Goldstein et al. 1995; Shriver et al. 1995; Slatkin 1995), and have been applied in a range of vertebrates (Bowcock et al. 1994; Meyer et al. 1995; Paszek et al. 1998; Richard and Thorpe 2001; Ayub et al. 2003; Mikul et al. 2007; Rout et al. 2008). Microsatellites were however only successful in reconstructing shallow phylogenies, enabling the study of evolutionary relationships between groups that have evolved independently for up to several million years. The main difficulty in reconstructing deeper phylogenies stemmed from restrictions imposed by microsatellite range constraints, which promote homoplasy (Noor et al. 2001; Estoup et al. 2002), irregularities and heterogeneity in the microsatellite mutation process (Ellegren 2004), and the degradation of the microsatellite over time (Buschiazzo and Gemmell 2006). To remedy this complication and increase the definition of deeper phylogenetic reconstruction using microsatellites, Martin et al. (2002) combined information from the repeat structure and the flanking sequences of a highly conserved microsatellite in sharks, while others relied solely on mutations occurring in flanking sequences (Makova et al. 2000; Domingo-Roura et al. 2005; Shepherd and Lambert 2005).

An altogether different approach to infer phylogenies is to use the binary information provided by the presence/absence states of characters in a set of species using models of mutation developed for morphological (discrete) or restriction site (binary)

characters. Equivalent or related strategies have been employed using genes (Huson and Steel 2004), indels (de Jong et al. 2003) and transposable elements (Bashir et al. 2005; Kriegs et al. 2006), but microsatellite-based tree reconstruction from presence/absence data have been unexplored to date.

Although all three resulting topologies compared relatively well with each other and with the current authoritative topology (Miller et al. 2007), important divergences existed regarding relationships within the Mammalia (e.g. Rodentia *vs*. Lagomorpha, Monotremata *vs*. Marsupilia). Overall, the MP topology was more accurate than Bayesian topologies, and almost perfectly identical to the authoritative topology, suggesting that, if correctly inferred, the mutational history of microsatellites along the mammalian phylogeny may be more successful than raw conserved microsatellite counts to infer interspecies relationships. Unfortunately, there is to date no theoretical development to test whether these models of mutation were in fact correctly inferring the retention history of microsatellites in vertebrate genomes. Besides, it is anticipated that a similar tree produced on the presence/absence of all microsatellites contained in each genome would produce a better tree, as there would not be any ascertainment bias stemming from the use of a single reference genome (here, human). Therefore, rather than seeking to provide an alternative to the currently authoritative vertebrate topology, our attempt to reconstruct an approximate phylogeny from binary microsatellite data aimed at testing whether there was a phylogenetic signal in the microsatellite conservation data using available algorithms.

In practice, models of mutation used for standard discrete (morphological) characters, as implemented in both Bayesian and MP phylogenetic reconstruction assume by default identical rates of change, i.e. 0→1 changes and 1→0 changes have identical frequencies. A

Bayesian approach proposing a directional morphological model of evolution was also implemented to allow the rates of change from 0→1 to differ from the rate of change from 1→0, but resulted in the most diverged topology (i.e. the m2p model). Clearly, models of evolution employed for morphological characters and restriction sites may arguably not be valid for microsatellite retention and, although it was opportune to assess this approach, future theoretical developments should endeavour to address this issue. Such development could help implement successfully phylogenetic and comparative analyses, not only tree reconstruction, but also ancestral character state reconstruction (Pagel et al. 2004).

An added spin-off of the present analysis was the successful demonstration that datasets of conserved microsatellites constructed from a restricted multiple genome alignment can be easily integrated into the comprehensive framework created from the 17-WA analysis. In theory, identifying human-platypus microsatellite pairs using the updated 28-WA would have been a more comprehensive alternative to the restricted 6-WA. In practice, handling the exceptionally large 28-WA files is hampered by memory limitations of conventional computational resources and current methods of sequence extraction.

3.6 Acknowledgments

A. Fouquet provided sound advice for phylogenetic reconstruction. We are also grateful to the team working at the UCSC Genome Browser for producing and providing the whole-genome alignment, and the team behind Galaxy for providing helpful computational tools, and for being prompt to provide both comments and advices when needed. V. Mencl also helped maintaining the stand-alone Galaxy version at the University of Canterbury.

Chapter 4

4 Design, optimization and implementation of degenerate comparative microsatellite primers for mammalian species

4.1 Abstract

Microsatellites are widely employed as genetic markers in various fields of research, but a number of pitfalls remain despite ongoing efforts to retain these sequences in the 21st century molecular ecologist's toolbox. In particular, isolating and developing *de novo* polymorphic microsatellites often requires expensive and intensive groundwork. It has been observed that microsatellites can be conserved in closely related species, and this property has since been exploited to transfer primer pairs designed in a focal species to amplify products in related non-focal species. However, in general, transferability decreases rapidly with increasing evolutionary distance, limiting the development of these cross-species markers. The recent surge of genome sequencing projects, especially for mammals, provides a new resource to develop primers for conserved microsatellite sequences for comparative analysis. In this chapter, I first describe how bioinformatic tools were used to identify microsatellites conserved across the Mammalia, and to design 19 wide-ranging primer pairs for comparative analyses from a random set of ~1000 conserved dinucleotide repeats. Second, I detail methods to optimize and implement these primers using a similar set of conditions, reducing both labour and human error. Third, I present results for nine genotyped loci and five sequenced loci in 18 species encompassing eutherian, metatherian and prototherian mammals. Finally, I evaluate these results and discuss how the methodology may be improved to help others wishing to develop comparative primers to amplify orthologous microsatellite loci in related mammalian species. Importantly, we anticipate that many more cross-species microsatellite markers (at least four times more, using the most stringent selection criteria) could be developed if genomewide conservation data were explored instead of the restricted random set used here.

4.2 Introduction

Microsatellites, or tandem repeats of short DNA motifs (1-6 bp), are hypermutable sequences, mutating at rates several orders of magnitude higher than the average genomic point mutation rate (Buschiazzo and Gemmell 2006). Typically, mutations in microsatellites derive from the addition or removal of one, though possibly several, motif(s). Due to the high frequency of these slippage events, inspection of microsatellite variability at a small number of loci and a large number of individuals may reveal unique haplotypes, i.e. multilocus genotypes, for all individuals. It is therefore possible to address issues such as discrimination, relationships, structure, and classification, not only at the population level (using allelic frequencies) but also at the individual level (using haplotypes). Such discriminatory power has promoted microsatellite loci as the molecular marker of choice for population genetic studies (Sunnucks 2000). Microsatellites have also been used successfully in forensic identification (Butler 2006), pedigree reconstruction and kinship assessment (Blouin 2003), phylogeography (Rossiter et al. 2007), shallow phylogeny reconstruction (Rout et al. 2008), linkage analysis (Park et al. 2008), gene hunting (*viz*. association studies, Tamiya et al. 2005), genome mapping (Luo et al. 2007), detection of selective sweep (Wiehe et al. 2007), genetic ecotoxicology (Yauk and Polyzos 2005) and epidemiology of infectious diseases (van Belkum 2007).

Such applications require scoring allele lengths (comprising the microsatellite sequence and sequences flanking the repeat array), either through direct sequencing of amplified fragments or through PCR product resolution on polyacrylamide gels, preferably using fluorescence-based methods (Ziegele et al. 1992), *viz*. fragment analysis or genotyping. The sizing accuracy, high sensitivity and reproducibility of fluorescence genotyping make this technique preferred over less effective radioactive

and non-radioactive staining (ethidium bromide and silver staining) methods (Ziegele et al. 1992).

A well-known caveat in using microsatellite genotyping is the impossibility to distinguish whether allele length variants are due to an addition/deletion of motifs in the repeat array, or to short indels in the flanking sequences (Estoup et al. 1995; Angers and Bernatchez 1997; Grimaldi and Crouau-Roy 1997). While direct sequencing is essential to understand the nature of variation among alleles, its routine use has been limited by its prohibitive costs, and the relatively low frequency and impact of these events. This has in turn encouraged the development and implementation of genotyping as an alternative cost-effective routine and fine proxy technique to detect microsatellite length mutations and exploit their extensive discriminatory power.

Accurate and reproducible scoring of microsatellites also depends strongly on the optimization of the PCR amplification of microsatellite products. This entails optimization of PCR reagent concentrations and cycling parameters, but also, and perhaps foremost, robust primer design.

Microsatellite markers are closely associated with their adjacent genomic region; when these flanking sequences are single-copy and conserved across individuals of the same species, they provide potential PCR priming sites for the specific amplification of orthologous loci across individuals. In addition, if these flanking sequences are further conserved in individuals of related species, they provide useful cross-species PCR priming sites (Moore et al. 1991; Schlötterer et al. 1991; FitzSimmons et al. 1995; Rico et al. 1996; Gemmell et al. 1997; Guillemaud et al. 2000; Kim et al. 2004; MacDonald et al. 2006). Microsatellites isolated from a single species (focal species) have been applied this way in population genetic studies of related species (non-focal species), yielding large amounts of genetic information

with little initial effort (Palo et al. 2001; Ruiz-Garcia 2005). In addition, comparative PCR primers can significantly increase the range and scope of applications of microsatellites, including: comparative genome mapping (Kondo et al. 1993; Varshney et al. 2005), species identification (Domingo-Roura 2002), inference of microsatellite evolution above the species level (Zhu et al. 2000), phylogenetic reconstruction (Martin et al. 2002), understanding mechanisms involved in speciation (Noor and Feder 2006), and community-based molecular ecology among multiple co-occurring species (Whitham et al. 2006).

Unfortunately, the success of microsatellite cross-species amplification rapidly decreases with evolutionary distance between focal and non-focal species (e.g. Primmer et al. 1996; Gemmell et al. 1997). This can be expected, considering that most microsatellites seem to evolve neutrally and are therefore only maintained by chance rather than active selection. (see Chapter 2). In addition, the accumulation of substitutions and indels in the flanking sequences and priming sites over time hinders the specificity and success of primer annealing to the target DNA sequence, resulting in so-called null alleles (Callen et al. 1993; Paetkau and Strobeck 1995). Given the intra- and intergenomic variability of mutation rates (Ellegren et al. 2003; Baer et al. 2007), the 'life expectancy' of priming sites, and thus the chances of successful cross-species amplification in related species are highly variable and locus-specific (Primmer et al. 2005). Barbara et al. (2007), in a comprehensive survey encompassing a total of 611 cross-species studies within three kingdoms, reviewed factors other than low phylogenetic distance that are positively correlated with transferability of microsatellites across species; among these factors, long generation time, mixed or outcrossing breeding systems, and high (source:target) genome size ratio were the most significant.

Even when these factors are taken into account to help identify transferable microsatellites, the standard cross-species PCR amplification approach (i.e. testing transferability of microsatellites isolated in focal species to non-focal species) still restrain the full potential of cross-species microsatellite markers. First and foremost, the possibility to transfer microsatellites in non-focal species obviously depends on the prior isolation of microsatellite markers in a closely related focal species, and therefore relies on substantial groundwork. In addition, most conserved microsatellite markers described to date are limited to closely related taxa, which dramatically limits the starting number of testable loci. Finally, lack of sequence knowledge in non-focal taxa implies that investigators are working blind and can only assume sequence conservation when designing and/or testing primers; therefore they can not reliably design the optimal (most conserved) primer pairs at any given locus. Consequently, reports of microsatellite transfers over a large evolutionary scale (above the genus level) have remained anecdotal (Rico et al. 1996; Fitzsimmons 1998; Moore et al. 1998), resulting more by chance rather than from a systematic and thorough approach. Recent advances in high-throughput sequencing techniques, sophisticated computational tools and the construction of comprehensive databases of putatively conserved sequences (e.g. EST and UniGene databases) have facilitated cross-species investigations (e.g. Stallings 1995; Farber and Medrano 2004; Liewlaksaneeyanawin et al. 2004; Perez et al. 2005; Varshney et al. 2005; Parida et al. 2006; Pashley et al. 2006). However, these approaches lack the magnitude and comprehensive scope that only large-scale whole-genome comparisons are able to provide.

In Chapters 2 and 3, I presented the first comprehensive surveys of microsatellites conserved in vertebrate genomes. This framework represents a unique opportunity to develop an extensive source of primers for microsatellite markers that work not only in closely, but also distantly, related mammalian and other vertebrate

species. To illustrate the full potential of this dataset in cross-species studies, I aimed to develop and optimize a standard protocol applied to a set of microsatellite primers conserved across the Mammalia, including eutherian, metatherian (marsupials) and prototherian (monotremes) species. The present chapter reveals a detailed description of the protocol used to select 19 appropriate conserved microsatellites, followed by the design, optimization and implementation of comparative primers for genotyping and sequencing purposes. Results are presented succinctly, and suggestions to improve the methodology are discussed. In addition, I provide brief guidelines that could be useful for others planning to use conserved microsatellites to develop comparative primers.

4.3 Materials and Methods

4.3.1 Collection of mammalian samples

A collection of DNA, blood or tissue samples from 20 unrelated individuals per species were brought together from generous donors around the world (Table 4.1). All donors confirmed that, to their best knowledge, all sample individuals were unrelated, although rats originated from inbred populations (Robertson and Gemmell 2004) and pilot whales were sampled from pod strandings (M. Oresmus, pers. comm.). Species choices were made to include sister species representatives for three of the four superorders of eutherians (Laurasatheria, Euarchontoglires and Afrotheria), marsupials and monotremes. Unfortunately, due to stricter restrictions on the export and/or use of xenarthran species, such as anteater and armadillo, we have not been able to source two species from this fourth eutherian superorder.

All subsequent steps were carried out at the University of Canterbury, except for the analysis of the chimpanzee samples for which all work was carried out at Arizona State University in Dr. Anne Stone's lab by Luz-Andrea Pfister. All work was carried out with prior agreement from all donors; in particular, work on pilot whale samples was carried out in compliance with the New Zealand Department of Conservation and local iwi requirements (no genetic modification).

4.3.2 Preparation of genomic DNA

When samples were not provided directly as DNA in solution, total DNA was extracted from either tissue (preserved at -80°C in ethanol or DMSO) or blood samples, using slight variations of the Chelex method (Walsh et al. 1991). Although cat blood was preserved in EDTA for less than 1 week prior to extraction, dog blood was stored at -80°C in EDTA for ~12 months before extraction.

For DNA extraction from blood, 3 µl of whole blood was added to 500 µl of sterile distilled water; tubes were left at room temperature for 30 min, occasionally mixed by inversion, and then centrifuged for 2 min at 15,000 g. The supernatant was carefully removed, leaving only 20-30 µl, and discarded. 5% Chelex was added to obtain a final volume of 100 µl. For DNA extraction from tissue, a 5-10 mm^3 piece of tissue was cut and placed into 100 µl of 5% Chelex in TE with 1 µl of proteinase K (20 mg/ml).

For both blood and tissue extraction methods, tubes were then vortexed for a few seconds, and placed into a shaking incubator at 58°C for 2 hours followed by incubation at 90°C for 8 minutes to denaturate the proteinase K. After this denaturing treatment, tubes were vortexed for a few seconds and centrifuged for 4 minutes at

20,800 g. The supernatant was transferred into new tubes, and kept at -20°C overnight before subsequent use.

All DNA extracts were quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop). An aliquot (~40 µl) of each DNA extract was diluted in TE to 20-50 ng/µl when necessary, and placed in 96-well plates for storage at -20°C between each use.

4.3.3 Identification of conserved mammalian microsatellites

Orthologous mammalian microsatellites were identified in the UCSC vertebrate 17-WA using a similar approach to that presented in Chapter 2. However, there were some significant exceptions in the methodology, including: (i) the analysis was limited to mammalian sequences; (ii) FASTA-formatted sequences were extracted in a pairwise fashion (human-other species) using Gmaj (http://globin.cse.psu.edu/dist/gmaj/); (iii) sequences were scanned with a modified version of Sputnik (La Rota et al. 2005), using the following parameters: –v 1 –u 5 –n -4 –s 8 –L 15 (motif length: 1-5 bp; mismatch penalty: -4; min score: 8, min array length: 15 bp).

Table 4.1: Nature and origin of mammalian samples

Superorder	Order	Common name	Scientific name	Seq*	CITES†	#‡	Type	Institution	Contact
Laurasiatheria	Cetartiodactyla	Cow	*Bos taurus*	Y	-	20	DNA	Lincoln U, New Zealand	J. Hickford
		Sheep	*Ovis aries*	Y	-	20	DNA	Lincoln U, NZ	J. Hickford
		Dolphin	*Tursiops aduncus*	N§	-	20	Tissue	Macquarie U, Australia	L. Moller
		Pilot whale	*Globicephala melas*	N	II	20	Tissue	Auckland U, NZ	M. Oremus, S. Baker
	Carnivora	Cat	*Felis catus*	Y	-	20	Blood	Gribbles Veterinary, NZ	-
		Dog	*Canis familiaris*	Y	-	20	Blood	U of Canterbury, NZ	I. Vargas-Jentzsch
	Eulipotyphla	Hedgehog	*Erinaceus europaeus*	Y	-	20	DNA	U of Canterbury, NZ	M. Hale
		Shrew	*Sorex araneus*	Y	-	20	Tissue	U of Lausanne, Switzerland	G. Yannic, J. Hausser
Euarchontoglires	Rodentia	Mouse	*Mus musculus*	Y	-	20	DNA	Köln U, Germany	D. Tautz
		Rat	*Rattus norvegicus*	Y	-	20	Tissue	U of Canterbury, NZ	B. Robertson
	Primates	Human	*Homo sapiens*	Y	-	20	DNA	National Cell Bank of Iran	A. Amanzadeh, F. Shokri
		Chimpanzee	*Pan troglytes*	Y	II	20	Outsourced	Arizona State U, USA	A. Stone
Afrotheria	Afrosoricida	Tenrec	*Echinops telfairi*	Y	-	20	Tissue	Field Museum, Chicago, USA	S. Goodman
	Sirenia	Dugong	*Dugong dugong*	N	I	20	DNA	James Cook U, Australia	A. MacMahon, D. Blair
Australidelphia	Diprotodontia	Tammar wallaby	*Macropus eugenii*	Y	-	16	DNA	Australian National U, Australia	J. Graves
	Dasyuridae	Quoll	*Dasyurus maculatus*	N	-	20	Tissue	U of New South Wales, Australia	M. Cardoso
n/a	Monotremata	Platypus	*Ornithorhynchus anatinus*	Y	-	20	DNA	U of Sydney, Australia	C. Whittington
		Echidna	*Tachyglossus aculeatus*	N	-	20	DNA/tissue	U of Tasmania, Australia	S. Nicol
								Kangaroo Island, Australia	P. Rismiller

* indicates whether genome sequence is partly or completely available to the public

† CITES Appendices for listed protected species (http://www.cites.org/eng/resources/species.html)

‡ Number of unrelated individuals sourced for the study

§ Sequencing of the genome of a related dolphin, *Tursiops truncatus*, is under way.

4.3.4 Conserved dinucleotide repeats

A subset of human dinucleotide microsatellites (length ≥14 bp) was isolated based on their broad conservation in mammalian species (present at least in five mammals, or in comparisons including at least human, either dog or mouse, and opossum); special care was taken to ensure that all orthologous microsatellites contained dinucleotide repeats. Using the interval position of microsatellites in the human genome, the conservation of flanking sequences (~250 bp either side) across mammals, including platypus, was reviewed by eye using the 28-way conservation track of the UCSC Genome Browser. Criteria for selection were: (i) presence of a microsatellite sequence in all species included in the present study, although exceptions were tolerated for low-coverage (2X) genomes of cat, armadillo, elephant and tenrec (because false negatives, i.e. sequence gaps, could not be reasonably ruled out), (ii) interspecies polymorphism (i.e. variable length of repeat array), (iii) at least ~20 contiguous base pairs on each side of the microsatellite perfectly or almost perfectly identical between comparison species, and (iv) total length of the potential amplicon not exceeding ~500 bp, as required for genotyping. Eventually, 73 alignments comprising a microsatellite and flanking sequences were pre-selected this way. These alignments were downloaded from the UCSC 28-way conservation track and converted to FASTA format using Galaxy (Giardine et al. 2005). Sequences from species included in the study were kept, together with those of armadillo (Xenarthra), elephant (Afrotheria) and opossum (Marsupilia), to produce a locus set that covered the breadth of the Mammalia. Where necessary, microsatellite flanking sequences were re-aligned manually using BioEdit (Hall 1999).

4.3.5 Comparative primer design

All alignments were submitted to PrimaClade (Gadberry et al. 2005); this web application runs Primer3 (Rozen and Skaletsky 2000) independently for each sequence, collating the results to identify comparative primers that bind across the alignment, while allowing for base degeneracy (Appendix, Table 4). A maximum of three degenerate sites per primer were allowed. Primers that overlapped gaps in the alignment were excluded, and only primers generating fragments smaller than 350 bp were kept for further selection. Using the java web-application NetPrimer and the developer's recommendations (PREMIER Biosoft International, http://www.premierbiosoft.com/netprimer/), potential primer pairs were tested for the presence of secondary structures (hairpins, self- and cross-dimerization), palindromes and repeats that could affect the amplification reaction through intra- and intermolecular interactions and non-specific annealing. Table 4.2 summarizes the overall set of unambiguous criteria that were applied to select comparative primers and increase the chances of successful amplification.

Following this selection process, the best possible primer pairs for 19 microsatellites were successfully designed and ordered from a commercial provider (Sigma) with an M13(-21) tail (5'-TGTAAAACGACGGCCAGT-3', Schuelke 2000) at the 5' end of one of the two primers (subsequently referred to as the forward primer). A list of all primers tested and their characteristics is displayed in Table 4.3.

In addition, I applied the same criteria to design primers for a locus containing the non-coding microsatellite with the widest range of conservation in mammals described to date, and located in the 3'-UTR of the NCAM1 gene (Moore et al. 1998).

Table 4.2: Selection criteria for designing comparative microsatellite primers

					Repeats		Stability of primer secondary structures (ΔG^*)					
$L_{expected}$	L_{primer} †	T_m ‡	ΔT_m	%GC§	2-6x	1x	3' HP	Int HP	3' SD	Int SD	3' CD	Int CD
<350	18-26	58-62	<1	30-62	<3	<6	>-2.00	>-3.00	>-5.00	>-6.00	>-5.00	>-6.00

$L_{expected}$: expected length of PCR products (bp); L_{primer}: primer length (bp); Tm: melting temperature (°C); ΔT_m: T_m difference between both primers; %GC: G+C content; 2-6x: number of tandemly repeated non-mononucleotide motifs (2-6 bp); 1x: length of mononucleotide runs; ΔG: Gibbs free energy required to break the secondary structure (kcal/mol); 3': 3'-end of primers; Int: Internal; HP: hairpin, SD: self-dimer, CD: cross-dimer.

4.3.6 Polymerase Chain Reaction (PCR)

Amplification conditions were optimized until only the DNA bands of the expected size were present as a single, or at least major, band observed on the electrophoresis gel. Optimal touch-down PCR (Don et al. 1991; Hecker and Roux 1996) conditions were found using a range of annealing temperatures, $MgCl_2$, final extension time, and primer, DNA and TMAC (Chevet et al. 1995) concentrations on 2 samples from 10 mammals, including tammar wallaby, platypus and echidnas.

Optimized PCRs were performed on a MasterCycler epGradient S (Eppendorf), in 15 µl of reaction volume containing 10 pmol of each primer, 2.5 mM $MgCl_2$, 200 mM each dNTP, 40 mM TMAC, 0.75 U of BioTaq DNA polymerase (Bioline), and 20-100 ng of genomic DNA template. The touch-down PCR cycling conditions included a hot-start step at 94°C for 3 minutes, followed by an initial annealing temperature T_{init} (generally 59°C, but see exceptions in Table 4.2), and a decrease of the annealing temperature at the rate of 2°C for every two PCR cycles (denaturation at 94°C for 15 s, annealing for 30 s and extension at 72°C for 20 s) until the target temperature ($T_{targ} = T_{init}$ - 10°C) was reached. We performed 26 regular

* Output from NetPrimer; criteria as recommended in the application's manual
† Preferentially 18-22 bp
‡ Output from PrimaClade
§ Preferentially 45-60% but lower values were tolerated for primers >22 bp

cycles at T_{targ} and samples were incubated at 72°C for 20 minutes for final extension. 3 µl of amplified products were loaded on 1.5 % agarose/TBE gel stained with BET, resolved by electrophoresis and visualized under ultraviolet (UV) light. Primer pairs resulting in multiple bands or no amplification in all or most species were discarded.

4.3.7 Microsatellite genotyping

PCRs were performed as described above; however, the reactions contained 5 pmol of forward primer and 10 pmol of M13 primer (1:2 ratio). Depending on signal intensity of bands under UV light, 0.5-2 µl of amplified product was combined with 10 µl of formamide and 0.3 µl of GeneScan 500LIZ size standard (Applied Biosystems), placed at 95°C for three minutes and in ice for 10 minutes. Fragment analysis was performed in an ABI3100 Genetic Analyzer (Applied Biosystems). Where possible, PCR products were pooled and run in groups of 2-4 markers per run using various fluorescent tags (VIC, NED, PET: Applied Biosystems; FAM: Sigma). Fragment sizes were scored with GeneMarker (Soft Genetics LLC).

4.3.8 DNA sequencing

We restricted this analysis to the five primer pairs that allowed successful genotyping in a broad range of species, namely C2-1218, C2-1915, C4-1514, C9-1918 and C17-4243 (Table 4.3). Four individuals per species per locus were selected for direct sequencing on a locus by locus basis based on homozygozity and, where possible, polymorphism.

Normal PCRs were performed first, and 3 µl of product was loaded in an electrophoresis gel to check for amplification success and presence of spurious bands. Because of the amount of primer dimers, two filter-plate purifications were necessary. The sequencing PCR was run using a standard protocol (Big Dye Terminator Cycle Sequencing Kit, Applied Biosystems), and products were prepared for sequencing in both directions in an ABI3100 Genetic Analyzer (Applied Biosystems) following the manufacturer's instructions. Sequences obtained for each locus were aligned with ClustalW (Thompson et al. 1994), and edited manually using BioEdit (Hall 1999).

4.4 Results

4.4.1 Quality of genomic DNA

Due to variance in tissue type, differences in preservation, and differences in the method of DNA extraction among samples from different species, the purity and quantification results were very variable between species and samples (data not shown). However, with the exception of DNA extractions from 4 dog blood samples, all estimates of DNA concentrations were higher than 20 ng/µl. The lower quality of dog DNA extracts probably stems from the relatively long storage of blood samples prior to extraction (~12 months).

4.4.2 Identification of microsatellites conserved across the Mammalia

Using a slight variation of the method described in Chapter 2, 126,306 human microsatellites were found to be conserved in at least one non-primate mammal, compared to 199,403 microsatellites identified in the search performed in Chapter 2 (63.34%). This discrepancy among searches is attributed to the different algorithm and built-in options implemented in Sputnik and SciRoKo. In particular, Sputnik only searches for repeated motifs of length 1-5 bp, and has a length cut-off of 15 bp, whereas SciRoKo looks for 1-6 bp repeated motifs with a length cut-off of 12 bp and therefore identifies comparatively more microsatellites. In addition, different purity parameters affect the final number of identified microsatellites. Overall, Sputnik is still a fairly good repeat finder and identified a workable amount of microsatellites in mammalian sequences; it was the best possible software solution available at the time of the analysis, but has been superseded recently by SciRoKo (Kofler et al. 2007).

Whereas the numbers of microsatellites identified differed significantly from results reported in Chapter 2, the relative proportions of human microsatellite conservation detected between species were similar using Sputnik and SciRoKo (Figure 4.1; Chapter 1, Figure 2.1A), confirming that the present dataset accounts well for the evolutionary conservation of microsatellites in mammalian genomes.

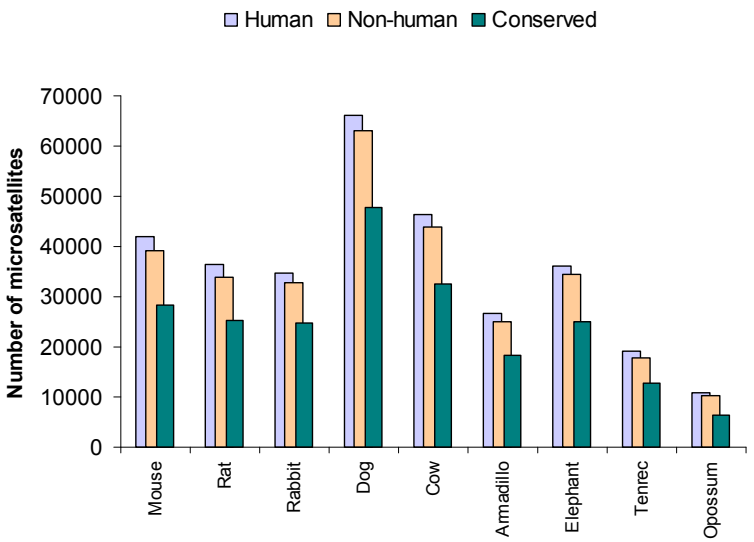


Figure 4.1: Conservation of human microsatellites in pairwise sequence alignments. Number of microsatellites identified in human sequences (blue), in non-human sequences aligned to the human genome (orange), and conserved between both species (green).

4.4.3 Broadly conserved dinucleotide repeats

Drawing on the identification of conserved microsatellites in mammalian genomes, we sought to isolate a large subset of potential microsatellite markers, and focused our search on dinucleotide repeats. Long dinucleotide repeats are often employed as genetic markers because of their abundance, ubiquity and mutability in vertebrate genomes (Tóth et al. 2000; Kelkar et al. 2008). We randomly isolated ~1000 human dinucleotide repeats covering all human autosomes and the X chromosome from a subset of broadly conserved microsatellites, i.e. (i) present in at least five of the nine non-primate species, or (ii) in human, opossum and either dog or mouse. Although the objective was to identify a subset of microsatellites conserved across all mammals, if such a set existed, these less stringent criteria of selection were preferred because (i) the sequence containing the microsatellites may be unavailable in one or more species; (ii) Sputnik has a length cut-off of 15 bp (effectively 14 bp for dinucleotide

repeats) and thus does not identify short microsatellites, which are more frequent in mammalian genomes (Dieringer and Schlötterer 2003); (iii) microsatellites may have disappeared in one or more species. Cases relating to the two latter points can help document the hypothesis of the microsatellite life cycle (Buschiazzo and Gemmell 2006), and were therefore considered when choosing selection criteria.

Of ~ 1000 broadly conserved dinucleotide microsatellite loci, 73 were selected by eye for the presence of highly conserved portions on either side of the repeat sequence for potential comparative primer design (Figure 4.2A). Although somewhat subjective and repetitive, this approach was both efficient and practicable for this limited number of starting microsatellites.

The proportion of microsatellites with potential comparative primer sites was fairly elevated (~7%) considering the breadth of the Mammalia and the expected accumulation of substitutions and indels in microsatellite flanking sequences. This proportion yet demonstrates that the broad retention of microsatellites across mammals does not necessarily occur in mutation-free regions.

Interestingly, a substantial number of polymorphic microsatellites (interspecies length variation) showed high sequence conservation on one side only, while the other side aligned poorly (Figure 4.2B). This observation could give support to the hypothesis that microsatellites are associated with recombination hotspots (Bagshaw et al. 2008) These variable yet conserved microsatellite loci may well point at unannotated putatively functional sequences that are selected for mutability (King and Kashi 2007). A thorough analysis might prove valuable.

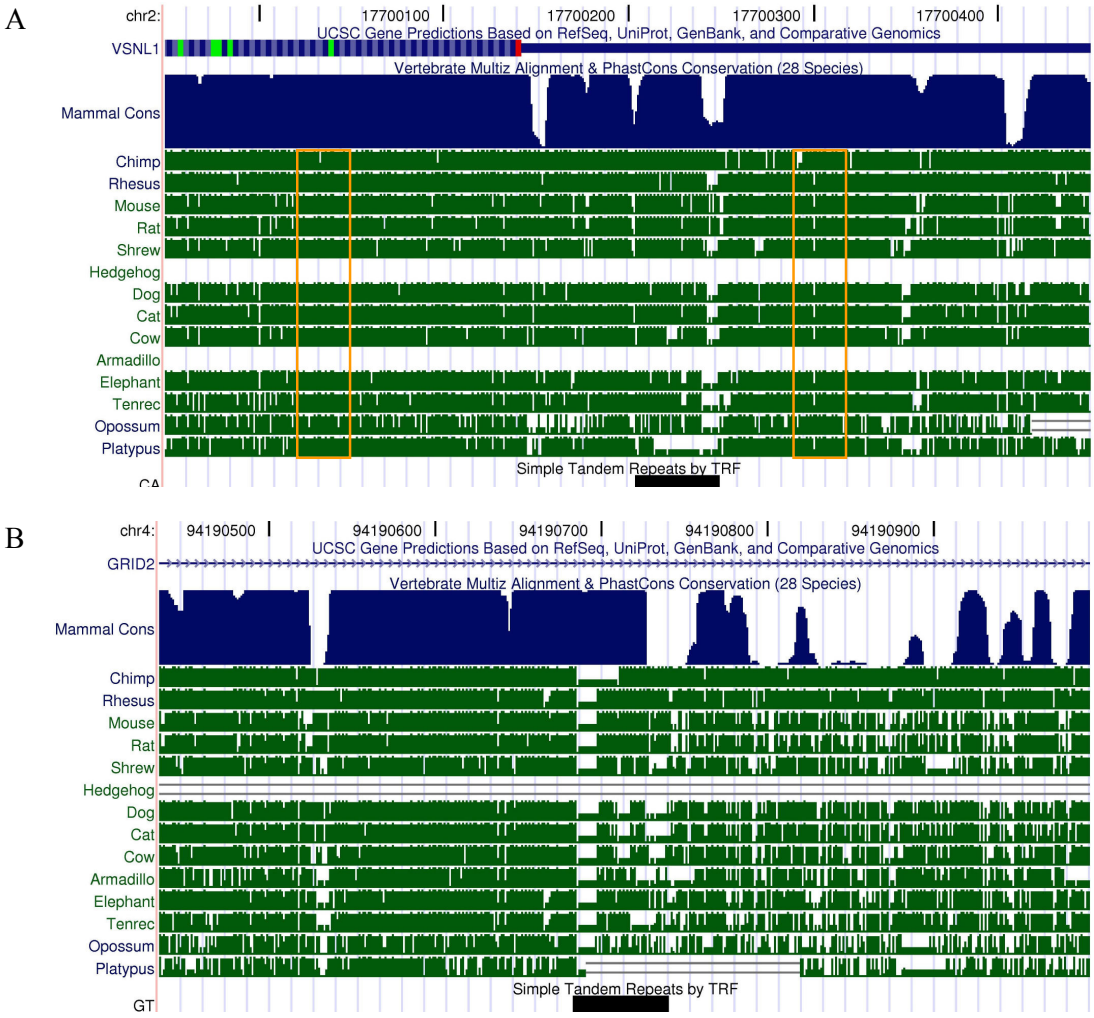


Figure 4.2: 28-way alignment of conserved microsatellites. (A) Both sides of the conserved microsatellite are conserved enough for potential primer design. Orange boxes indicate comparative primer positions. Hg18:chr2:17,699,950-17,700,450. (B) Only the sequence upstream of the microsatellite is well conserved, whereas the downstream sequence is very divergent. Hg18:chr4:94,190,435-94,190,994. Images were captured from the UCSC Genome Browser.

4.4.4 Comparative primer design

Primers were successfully designed for 19 of the 73 pre-selected loci, i.e. ~2% of our initial random subset of dinucleotide microsatellites widely conserved in the Mammalia. None of the 19 loci were suitable to design non-degenerate primer pairs. Table 4.3 shows the characteristics of the best possible primer pairs for each locus, as

well as the expected length of the product they would amplify in species used in this study for which public sequence information was available.

All 19 loci showed length polymorphism between species. Ideally, a cross-species genetic marker would contain a polymorphic microsatellite and indel-free flanking sequences (i.e. without any gap in the alignment); thus, any length variation of amplified products would be safely attributable to additions or deletions of motifs in the microsatellite sequence. In theory, this is not impossible, as highly conserved sequences in vertebrates (Woolfe et al. 2005) and indel-purified sequences in mammals (Lunter et al. 2006) have been described. In practice, whereas most length differences could be attributable to mutations in microsatellite sequences, indels also occurred in the flanking sequences and none of the 19 alignments were indel-free (e.g. Figures 4.3-7). These observations confirmed previous findings that sequence information is essential to understand interspecies length variation at microsatellite loci (Primmer and Ellegren 1998). In addition, other mutable short tandem repeats, if present in the amplified sequence, might bias interpretation of allele length variation, e.g. mononucleotide runs (see tenrec sequence at locus C2-6868, Figure 4.5B), which are the most mutable type of microsatellite identified in the human and chimpanzee genomes (Kelkar et al. 2008), and this pattern may well extend to other species.

The likelihood of finding indels in sequences did not seem to depend on the time of divergence from human, e.g. at the C2-1218 locus, many indels were found in human-chimp comparisons (Figure 4.3, cf. positions 1, 4, 11, 30, 34, 48 and 50). In contrast, the frequency of base substitutions increased with increasing distance from human, in agreement with the neutral model of evolution, e.g. opossum and platypus sequences showed significantly more substitutions than eutherian sequences. This observation confirmed that the chances of designing robust comparative primers decrease with increasing evolutionary distance.

4.4.5 Structural and functional aspects of selected conserved microsatellites

The interspecies stability of the microsatellite internal structure (motif usage, purity, repeat array length, complexity) was very variable between the 19 selected loci, as illustrated in Figures 4.3-7. Such structural changes should be considered when interpretations regarding allele length changes are made (Chapter 5). For example, the C2-1218 locus (Figure 4.3) contains a highly conserved simple dinucleotide $(CA)_n$ repeat that has accumulated few point mutations throughout mammalian evolution, although exceptions occurred in mouse, shrew and opossum. Length variation of the microsatellite sequence for this locus can mostly be attributed to addition/deletion of (CA) motifs. In contrast, the C4-1514 (Figure 4.6) locus contains a (CA)-based repeat where many lineage-specific point mutation events took place, altering the initial simple structure as seen in platypus into a complex amalgam of short microsatellite sub-units separated by degenerated motifs.

The variable stability of repeat structures might be partly explained by selective constraints acting on these conserved, thus putatively functional, microsatellites. Interestingly, Riley and Krieger (2004; 2005) observed repeat replacements at orthologous microsatellites situated in UTRs and introns of genes of similar functions. Riley and others later extended these results and demonstrated that some of these repeat replacements in UTRs still preserved folding potential, thus suggesting repeat selection at the level of higher order functional structure rather than primary structure (Riley et al. 2007). In 18 loci, no such replacement was observed, although short expansions of motifs derived from the ancestral motif occurred, generating compound microsatellites. One significant exception was however observed at the CX-4344 locus, where a $(GA)_n$ microsatellite located in the 5'-UTR of

the human NDP gene (Table 4.3), which was highly conserved across mammals (and also found in chicken and lizard, but is degenerate in frog), was replaced almost entirely by a $(CA)_n$ repeat in platypus (data not shown). Such motif replacements are intriguing considering the otherwise high conservation of these microsatellite loci across mammals, and a comprehensive assessment of these events should be considered elsewhere (Chapter 5).

The location of the other 18 microsatellites in introns, UTRs or IGRs of the human genome is shown in Table 4.3, as well as the name of the genes containing these microsatellites (or the closest gene in the case of microsatellites located in IGRs). No significant difference appeared as all three locations were well represented, with 7 selected loci located in UTRs (5 in 3'-UTRs and 2 in 5'-UTRs), 7 in introns, and 5 in IGRs. However, this also indicated that fewer non-genic microsatellites were suitable to develop comparative primers compared to genic microsatellites, which is not surprising because non-genic sequences accumulate more substitutions than actively transcribed regions (Ellegren et al. 2003). Many of these highly conserved microsatellites were associated with genes of similar and essential functions: five genes encoding zinc finger proteins (ZNF238, ZEB2, ZBTB20, ZNF608 and ZNF 536), three genes encoding transcription factors (LCORL, PBX3 and FOXG1B), and two genes encoding homeobox proteins (MEIS1 and HOXB3). Other microsatellites were also associated with genes involved in primary functions: neuronal intracellular signalling (VSLN1), neuronal cell adhesion (NCAM1), development of nervous system (NRN1), skeletal development (GCF5), regulation of apoptosis (BMF). Finally, one microsatellite was situated near a candidate gene for Schizophrenia (MRDS1), and two were associated with clones of unknown functions.

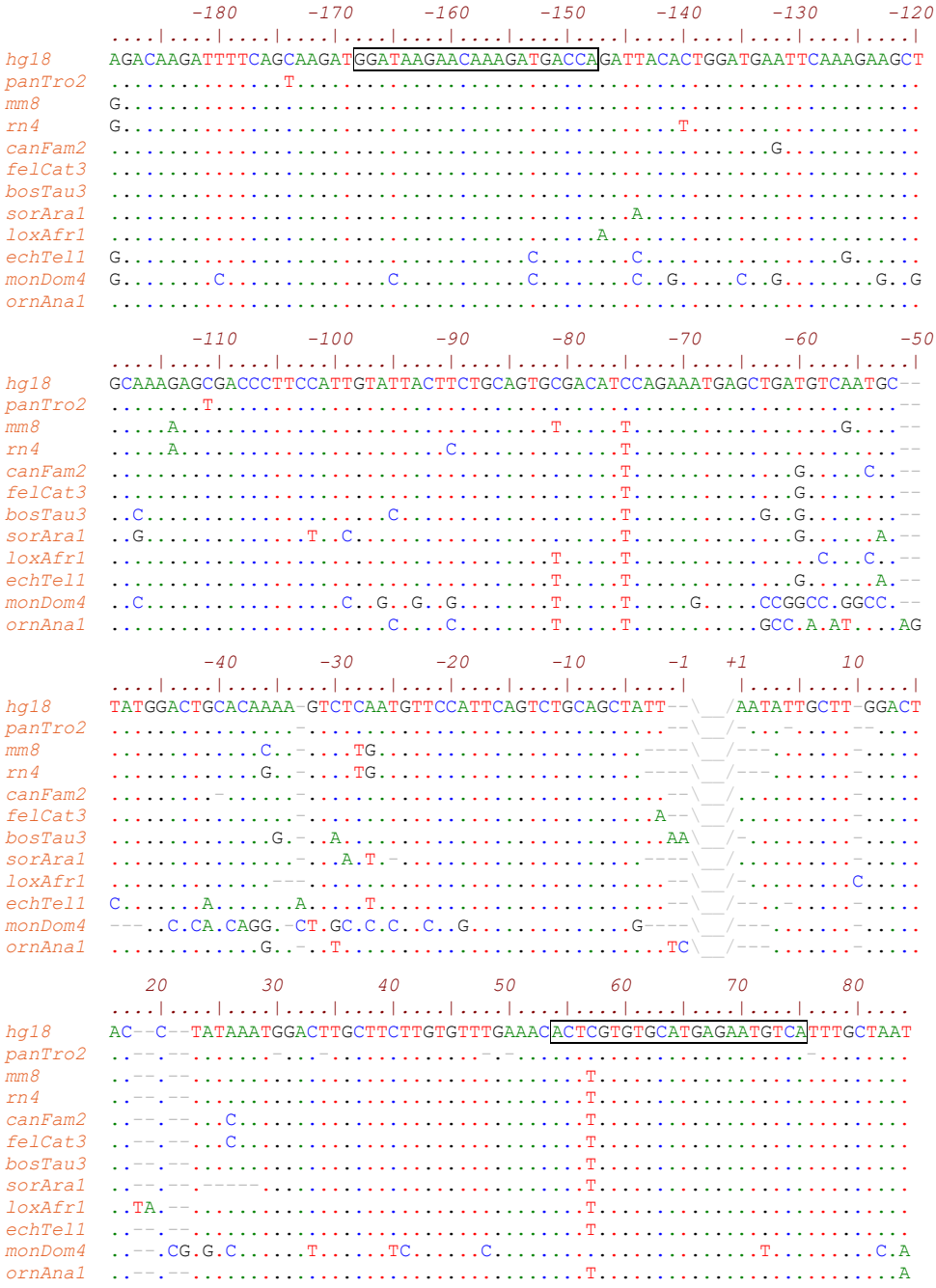
Notes for Table 4.3:

F: forward primer; R: reverse primer; L_{prim}: Primer length; DS: number of secondary (W, S, M, K, R and Y) and tertiary (B, D, H and V) degenerate sites; %GC: G+C content; T_m: melting temperature; T_{init}: initial temperature (see methods); *Hsa*: human, Ptr: chimpanzee; *Mmu*: mouse, *Rno*: rat, *Cfa*: dog, *Fca*: cat, *Bta*: cow, *Sar*: shrew, *Eeu*: hedgehog, *Ete*: tenrec, *Meu*: tammar wallaby, *Oan*: platypus; Genotype: indicates if the locus has been genotyped; Sequence: indicates if the locus has been sequenced. Expected lengths were calculated based on the 17-WA, or directly from trace files using MegaBLAST (http://www.ncbi.nlm.nih.gov/BLAST/mmtrace.shtml, accessed 15/03/08) when applicable, i.e. when the sequence is absent in the 17-WA and for tammar wallaby, which is not included in the 17-WA.

Table 4.3 : Characteristics of 19 primer pairs selected for optimization. See notes on opposite page

Name	Sequence	L_{prim}	DS 2nd	DS 3rd	%GC	T_m	T_{init}	Expected fragment length (M13 excluded) *Hsa*	*Ptr*	*Mmu*	*Rno*	*Cfa*	*Fca*	*Bta*	*Sar*	*Eeu*	*Ete*	*Meu*	*Oan*	Location (Gene)	Genotype	Seq
C1-2125	F=TCAGRGACTGGACCTTAGAAGA	22	1	-	50.00	60.24	59	133	133	143	137	131	145	133	139	135	?	?	276	3'-UTR (ZNF238)		
	R=AGRCTGCTTTTAGACCGTACC	21	1	-	52.38	59.80																
C2-1218	F=GGAYAAGAACAAAGAYGACCA	21	2	-	47.62	60.14	60	281	276	295	285	272	273	267	316	**266**	266	?	248	3'-UTR (VSLN1)	Y	Y
	R=TGAMATTCTCATGCACACRAGT	19	2	-	40.91	59.22																
C2-6868	F=CTTCCTCCAGAAGCCTCCTT	20	-	-	55.00	59.95	59	233	233	273	240	**264**	?	233	255	234	276	?	357	3'-UTR (MEIS1)	Y	Y
	R=TGGTGRGTAAATGAATTAGACATGC	25	1	-	36.00	59.47																
C2-1915	F=TCTTCTTCTTAATCACAATTCAGYC	25	1	-	36.00	60.14	59	174	160	197	176	184	?	173	228	172	?	**201**	223	Intron (ZEB2)	Y	
	R=CAAACAGCSMCAAACACC	18	2	-	55.56	59.94																
C3-1615	F=TGCTGTCTGGTAGRAGGGCT	20	1	-	55.00	60.01	59	233	231	236	235	244	244	233	233	223	221	?	?	Intron (ZBTB20)		
	R=TTTCCTGCTCYYGATGTT	18	2	-	50.00	59.17																
C4-1514	F=GGCATGTAWGTGGTTTTTGAACT	23	1	-	39.13	60.16	58	282	283	314	275	291	304	278	**322**	322	**280**	?	225	Intron (LCORL)	Y	
	R=GATCTGGAACACTGAAGACKC	21	1	-	52.38	59.87																
C5-1211	F=ACKTGGCAGCAGACCCTGT	19	1	-	57.89	60.88	59	269	269	280	280	258	**266**	260	259	261	263	?	260	3'-UTR (ZNF608)		
	R=TGGATCACATGGRGAAGC	18	1	-	55.56	60.43																
C6-5453	F=TTAGTMAAACCATTTGCGACY	21	2	-	42.86	59.99	59	275	271	261	**259**	266	273	**252**	?	268	?	**244**	273	5'-UTR (NRN1)		
	R=TGGGCGTGRGGCTGWRCT	18	3	-	61.11	60.43																
C6-1112	F=AATTGCTGYGCTAATTACAACAT	23	1	-	34.78	60.06	59	153	151	160	162	151	186	147	?	155	144	**150**	146	IGR (MRDS1)	Y	
	R=GACTTCTCCCAWGGCATVA	19	1	1	52.63	59.57																
C9-1918	F=GCCTTGCCAYGACCAYTT	18	2	-	50.00	59.46	59	302	313	303	296	301	**330**	307	?	311	323	?	293	IGR (PBX3)	Y	
	R=GCTCYSCGCTCAATTTAATT	20	2	-	45.00	60.41																
C11-1417	F=GGCAARAGACACTAGAGVARTG	22	2	1	50.00	59.15	59	207	218	228	225	179	203	197	?	204	184	**228**	?	3'-UTR (NCAM1)		
	R=RGYCMAGACGGTCCCCAA	18	3	-	55.56	59.02																
C13-7574	F=AATCAATATGGAAGGAGCGGT	21	-	-	45.00	59.81	59	253	255	275	303	254	257	254	253	247	267	**327**	265	Intron (BX647243)		
	R=HGCYTTTGTTTCAAGATGGT	20	1	1	42.86	60.11																
C14-2527	F=GGATGCTGAAATGACAGGC	19	-	-	52.63	59.19	59	249	251	236	266	262	?	267	245	?	294	?	?	IGR (FOXG1B)		
	R=GGTCTTCACACCAGGGRA	18	1	-	61.11	59.42																
C14-9692	F=TTAAGTGATTTTTGTAATGGTCCG	24	-	-	33.33	59.35	59	234	227	234	239	215	**222**	209	?	?	239	**205**	210	IGR (BC043585)	Y	
	R=YACCTCCAACAACTTCCTGTAAT	23	1	-	43.48	59.44																
C15-3531	F=CCAAGAGGGRTTATTTATGGC	21	1	-	47.62	59.69	59	228	226	298	224	240	256	244	265	223	?	**239**	195	Intron (BMF)	Y	
	R=TGATTTGACACTGKGARGG	19	2	-	52.63	59.46																
C17-4243	F=TGCMKAGGTTGCAGAATA	18	2	-	50.00	58.89	59	311	313	311	317	307	?	306	314	**305**	325	?	299	Intron (HOXB3)	Y	
	R=TRTCACATTTTACGACAATARCATTA	26	2	-	30.77	59.87																
C19-3338	F=GCTGTCCACCRAGATTCACC	20	1	-	55.00	60.52	59	206	204	197	232	260	?	**196**	226	?	197	**170**	?	IGR (ZNF536)		
	R=GCATYCCTGCCTTTRACTAA	20	2	-	50.00	60.21																
C20-3430	F=GGAAGAATGGCGTAATGCTG	20	-	-	50.00	60.61	59	185	167	201	**177**	180	**173**	162	207	184	**177**	**211**	?	Intron (GDF5)		
	R=CGTTTCCAATTCCYGAGTT	19	1	-	47.37	59.92																
CX-4344	F=GCCTGATTGATATATGMCTGS	21	2	-	47.62	59.55	59	212	212	216	210	202	196	204	204	226	241	?	214	5'UTR (NDP)		
	R=AAAGAGGTCSYGYTACTTCCA	21	3	-	47.62	59.98																

A

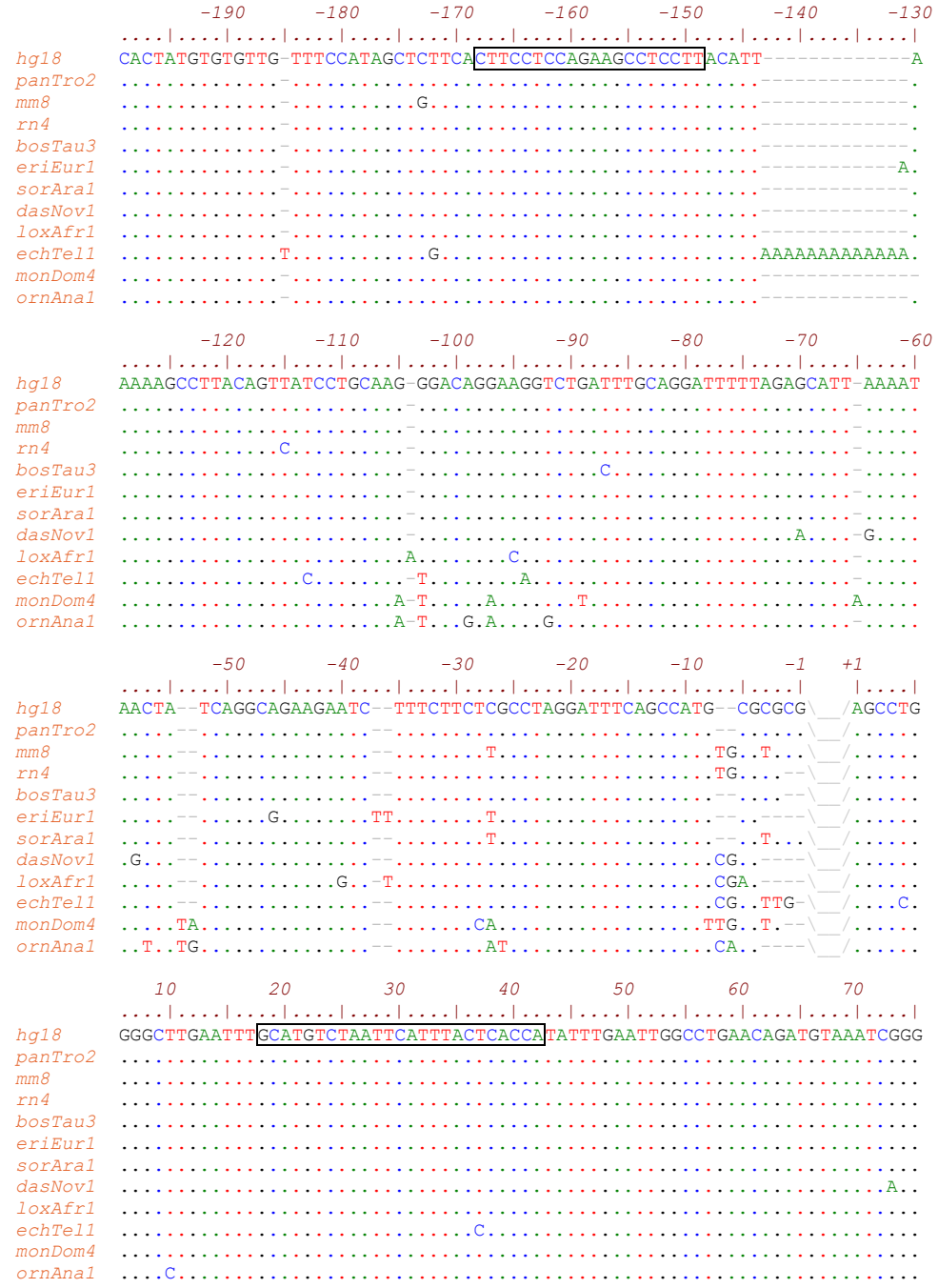


B

hg18	(46)	\(CA)$_{23}$ /
panTro2	(49)	\(CA)$_{24}$/
mm8	(65)	\(CA)$_{15}$(CCACACC)$_4$C(CA)$_3$/
rn4	(53)	\(CA)$_{20}$CTA(CA)$_6$ /
canFam2	(38)	\(CA)$_4$TA(CA)$_{10}$CG(CA)$_3$/
felCat3	(38)	\(CA)$_{19}$/
bosTau3	(31)	\(CA)$_{15}$G/
sorAra1	(89)	\(AC)$_{19}$AT(AC)$_5$GC(AC)$_5$ATGC(AC)$_2$A(AC)$_4$AT(AC)$_2$(AT)$_2$/
loxAfr1	(34)	\(CA)$_4$TA(CA)$_4$TA(CA)$_7$/
echTel1	(33)	\(CA)$_{16}$C/
monDom4	(39)	\(CA)$_2$CTCG(CCCACA)$_2$(CTCACA)$_2$(CA)$_2$CCA/
ornAna1	(12)	\(CA)$_4$(A)$_4$/

Figure 4.3: UCSC 28-way alignment of the C2-1218 locus showing species of interest. (A) Flanking sequences. Underscores represent the microsatellite sequence; positions are counted upstream and downstream from the microsatellite. Boxes indicate primer sites, dashes gaps and dots bases identical to human. (B) Microsatellite sequence. Array length is shown in brackets. UCSC assemblies: Human (hg18), chimp (panTro2), mouse (mm8), rat (rn4), cow (bosTau3), dog (canFam2), cat (felCat3), shrew (sorAra1), elephant (loxAfr1), tenrec (echTel1), opossum (monDom4), platypus (ornAna1).

A



B

hg18 (42) \(CT)$_5$TT(CT)$_4$(T)$_4$C(CT)$_4$CC(CT)$_2$TTCT/

panTro2 (42) \(CT)$_5$TT(CT)$_4$(T)$_4$C(CT)$_4$CC(CT)$_2$TTCT/

mm8 (80) \(CT)$_{29}$T(CTC)$_2$C(CT)$_2$CC(CT)$_2$TTCT/

rn4 (50) \(CT)$_7$CC(CT)$_6$(TC)$_2$(CT)$_4$CC(CT)$_4$/

bosTau3 (42) \(CT)$_5$TT(CT)$_5$(T)$_4$(CT)$_4$CC(CT)$_4$/

eriEur1 (40) \(CT)$_5$AT(CT)$_4$(T)$_3$C(CT)$_5$CC(CT)$_4$/

sorAra1 (64) \(CT)$_3$TT(CT)$_5$TT(CT)$_4$CCTT(CT)$_5$TCTT(CT)$_4$CC(CT)$_4$/

dasNov1 (42) \(CT)$_5$TT(CT)$_4$(T)$_3$C(CT)$_4$CC(CT)$_4$/

loxAfr1 (42) \(CT)$_5$TT(CT)$_5$TTC(CT)$_4$CC(CT)$_4$/

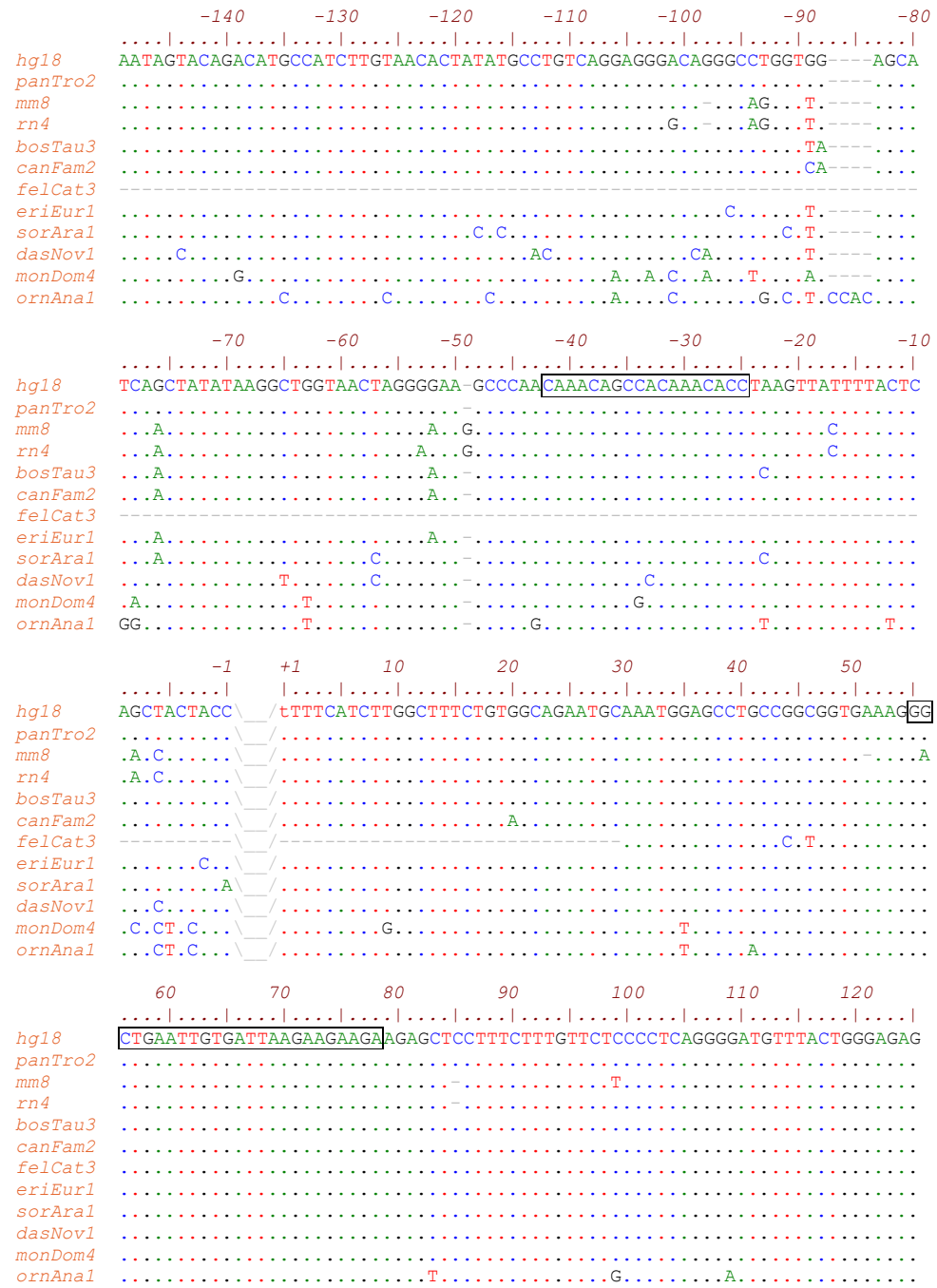
echTel1 (71) \(CT)$_{22}$T(CT)$_3$TC(CT)$_4$CC(CT)$_2$TTCT/

monDom4 (42) \(CT)$_{11}$TT(CT)$_4$CCTT(CT)$_3$/

ornAna1 (166) \(CT)$_3$TGCTCATCTTTCTGGCCTGT(CT)$_2$(C)$_5$(CT)$_2$GTC(T)$_4$G(CT)$_2$(CCCT)$_2$(CT)$_4$\\
\\CC(CT)$_5$(C)$_8$TTACA(CT)$_6$AC(CT)$_3$CCT(C)$_4$ATGACTGTTCCCCTCCCATTTCT\\
\\(C)$_9$AT(CT)$_4$TCCT/

Figure 4.4: Alignment of the C2-6868 locus. UCSC assemblies: hedgehog (eriEur1), armadillo (dasNov1). See legend Figure 4.3.

A



B

hg18 (51) \CTGTGC(CT)$_{15}$TT(TC)$_5$TTC/

panTro2 (37) \CTGTGC(CT)$_{14}$TTCT/

mm8 (75) \(CT)$_2$GAGC(CT)$_2$CC(CT)$_{14}$GT(CT)$_2$GT(CT)$_{12}$C/

rn4 (53) \(CT)$_2$GC(CT)$_{23}$C/

bosTau3 (50) \(CT)$_2$GC(CT)$_2$T(CT)$_7$(C)$_4$(CT)$_9$TTC/

canFam2 (61) \(CT)$_3$(CCT)$_2$(CTCTCCT)$_2$(CT)$_{16}$TTC/

felCat3 –

eriEur1 (50) \(CT)$_2$GCCTGCTT(CT)$_{15}$TTCTTTC/

sorAra1 (105) \CAT(TC)$_2$TG(TC)$_2$TG(CCTCT)$_2$GC(CTCTGT)$_2$(CT)$_4$(GTCTCT)$_2$(CTGTCT)$_2$\\
\\(CT)$_7$(C)$_3$(T)$_3$(C)$_3$(TC)$_2$CC(TC)$_3$TTTC/

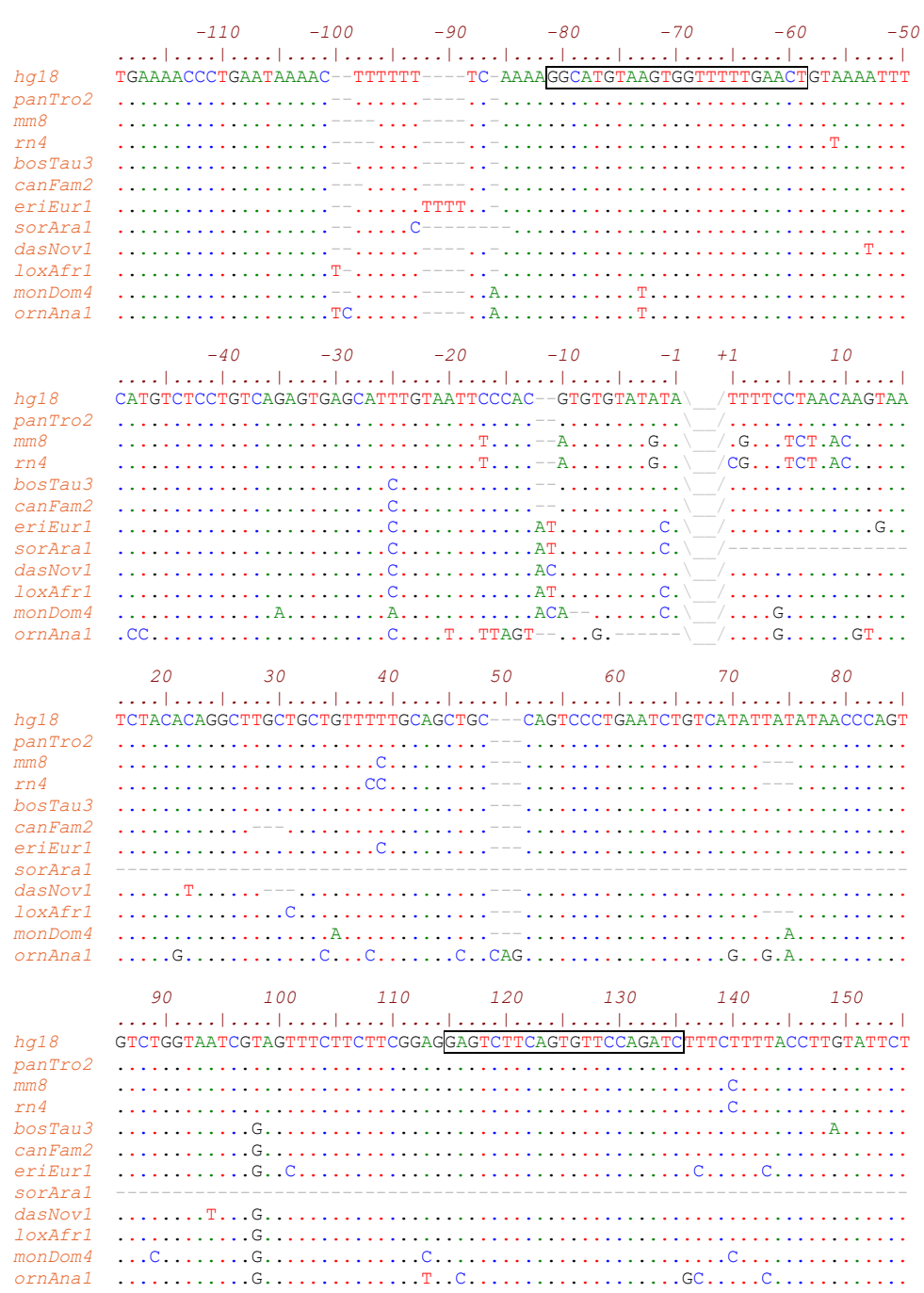
dasNov1 (67) \(CT)$_2$GCCTT(CT)$_9$GT(CT)$_3$GT(CT)$_5$TT(CT)$_8$TTC/

monDom4 (153) \(CT)$_2$GCT(C)$_6$ACC(T)$_3$(CT)$_{19}$CA(CT)$_{44}$A(T)$_4$/

ornAna1 (100) \ CTGTTC(T)$_6$C(A)$_5$(T)$_5$ATC(CCT)$_7$(CTT)$_2$T(CT)$_2$T(CT)$_2$CC(CT)$_4$TT(CT)$_5$(T)3A(T)$_7$CCTC /

Figure 4.5: Alignment of the C2-1915 locus. See legend Figures 4.3 and 4.4.

A



B

hg18 (70) \GA(CA)$_3$AATA(CA)$_2$CC(TG)$_2$(CA)$_2$TA(CA)$_4$CGC(A)$_3$(CA)$_9$CT(CA)$_2$AACA/

panTro2 (71) \GA(CA)$_3$AATA(CA)$_2$CC(TG)$_2$(CA)$_2$TA(CA)$_4$GC(A)$_3$(CA)$_9$C(CA)$_3$AACA/

mm8 (104) \GATA(CA)$_5$AAGACT(CA)$_2$TA(C)$_4$TGTA(CA)$_2$TACATTCCT(GCAC)$_2$A(CG)$_3$(CA)$_4$(CT)$_2$(CA)$_{10}$\\
\\GTCTCG(CA)$_2$/

rn4 (66) \GATA(CA)$_6$AAGACT(CA)$_3$(C)$_3$TTGTA(CA)$_3$CC(CA)$_5$CC(CA)$_5$/

bosTau3 (66) \GA(CA)$_3$(A)$_4$(CA)$_2$CC(TG)$_2$(CA)$_2$TA(CA)$_2$AA(CA)$_3$TACACGCTCACG(CA)$_2$TACA(C)$_3$G/

canFam2 (82) \GA(CA)$_3$AA(CA)$_3$CC(TG)$_2$(CA)$_2$TA(CA)$_6$CGCCAACG(CA)$_2$TA(CA)$_3$CG(CA)$_2$TA(CA)$_5$AACA/

eriEur1 (107) \(CA)$_4$(A)$_3$TA(CA)$_3$CC(T)$_3$(GC)$_2$AC(GC)$_2$GT(GC)$_3$A(CA)$_2$TA(CA)$_{12}$TATGC(A)$_3$(CA)$_7$TA(CA)$_3$AATA/

sorAra1 -

dasNov1 (48) \GACACGC(A)$_3$(CA)$_3$CC(TG)$_2$(CA)$_2$TA(CA)$_4$CG(CA)$_3$TACA/

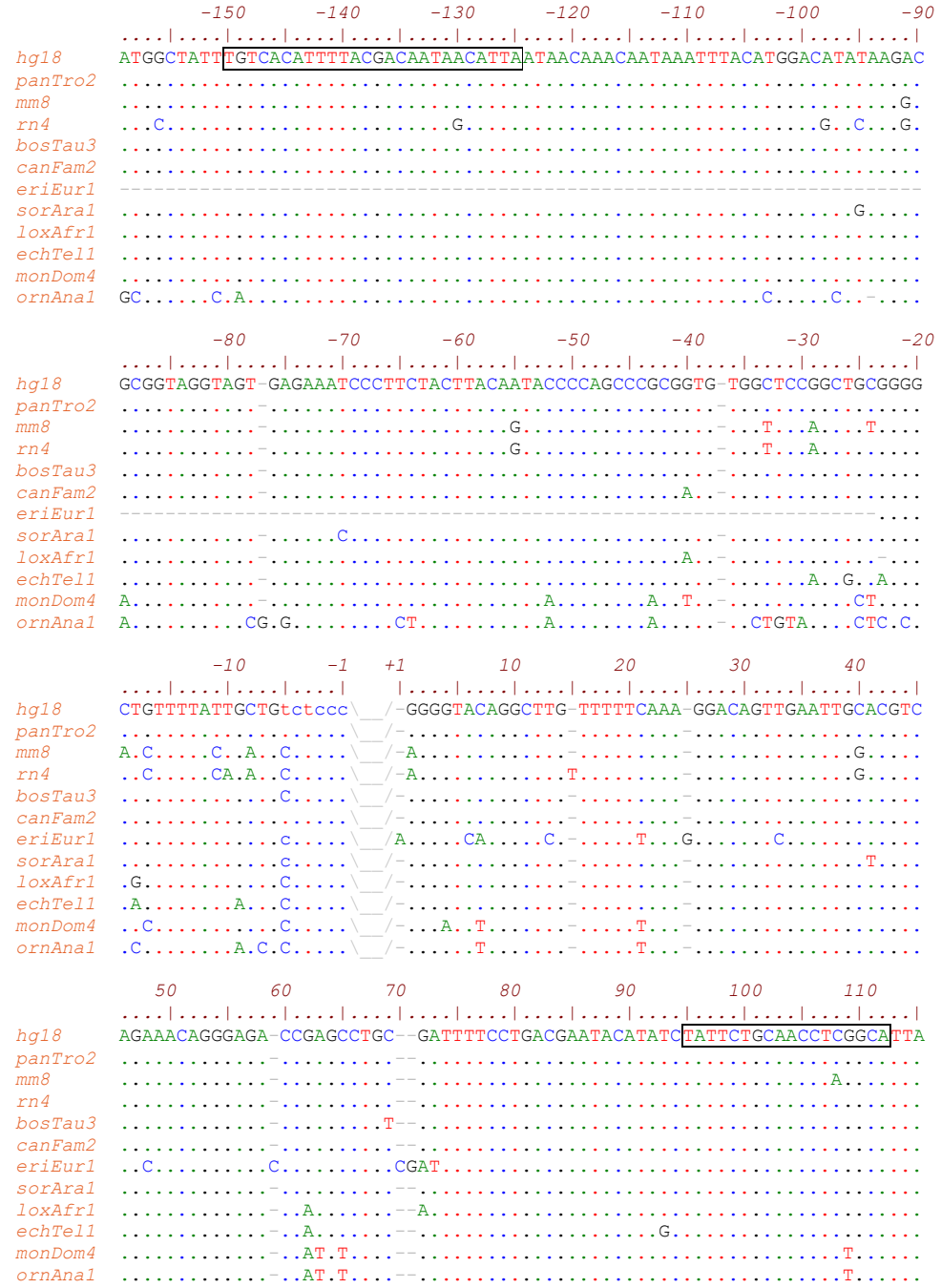
loxAfr1 (64) \GA(CA)$_3$AATA(CA)$_2$CC(TG)$_2$(CA)$_2$TA(CA)$_4$CGC(A)$_3$(CA)$_6$(CG)$_2$C(A)$_3$CA/

monDom4 (75) \GA(CATACA)$_2$T(C)$_3$AGTGCG(CA)$_2$CG(CA)$_5$CG(CA)$_2$CT(CA)$_3$(TA)$_2$(CAAA)$_2$(CA)$_3$TCA/

ornAna1 (16) \(C)$_{10}$(A)$_4$CA/

Figure 4.6: Alignment of the C4-1514 locus. See legend Figures 4.3 and 4.4.

A



B

hg18 (55) \(TTTC)$_2$(TC)$_7$CC(TC)$_5$(C)$_5$(TC)$_2$(T)$_8$(TC)$_2$/

panTro2 (57) \(TTTC)$_2$(TC)$_7$CC(TC)$_5$(C)$_4$(TC)$_3$(T)$_9$(TC)$_2$/

mm8 (55) \(TTTC)$_2$(TC)$_{15}$CC(TC)$_3$(T)$_5$(TC)$_2$/

rn4 (60) \TT(TC)$_{19}$CC(TC)$_2$(T)$_{10}$CCTC

bosTau3 (50) \(TC)$_2$(T)$_4$(TC)$_6$CC(TC)$_4$(C)$_4$(TC)$_2$(T)$_8$(TC)$_2$/

canFam2 (51) \(TC)$_2$(T)$_4$(TC)$_7$CC(TC)$_3$(C)$_4$(TC)$_3$(T)$_7$(TC)$_2$/

eriEur1 (52) \(TTC)$_2$TA(TCC)$_2$T(TCC)$_2$CCTT(TCT)$_2$(C)$_5$(TCT)$_2$(T)$_5$(TC)$_2$/

sorAra1 (58) \(TC)$_2$TT(TC)$_9$CC(TC)$_6$(C)$_4$(TC)$_2$(T)$_8$(TC)$_2$/

loxAfr1 (54) \(TC)$_2$(T)$_4$(TC)$_5$CC(TC)$_5$(C)$_4$(TC)$_3$(T)$_{10}$(TC)$_2$/

echTel1 (68) \TGTCTT(TC)$_8$CC(TC)$_3$TT(TC)$_2$(C)$_4$(TC)$_3$(C)$_3$(TC)$_6$(T)$_5$(TC)$_2$/

monDom4 (55) \TCTT(TC)$_9$TT(TC)$_5$(C)$_4$(TC)$_4$(T)$_5$(TC)$_2$/

ornAna1 (42) \(TC)$_2$(T)$_4$(TC)$_4$CCTTGC(TC)$_2$CC(TC)$_3$(T)$_5$TCTT/

Figure 4.7: Alignment of the C17-4243 locus. See legend Figures 4.3 and 4.4.

4.4.6 PCR optimization

The commonly preferred method for genotyping microsatellites is to synthesize one of the primers with a fluorescent label at the 5' end; amplified products are consequently labelled and will be detectable on a DNA sequencer, and their length compared to a size standard for scoring. However, fluorescent primers have a low output:cost ratio when many primer pairs require testing and the analysis involves a limited number of individuals. For these reasons, M13-genotyping was chosen as a cost-effective alternative in this study. According to the published method (Schuelke 2000), a fluorescently-labelled M13 primer (5'-TGTAAAACGACGGCCAGT-3') is added to the PCR mix, incorporating products amplified from standard primers after a few cycles and subsequently acting as the main forward primer; in principle, most amplified products are consequently labelled. This approach requires the optimization of primer concentrations using a range of (forward primer:M13-primer) concentration ratios. Excessive amounts of M13 primer could increase primer dimers and/or inhibit the first stages of the PCR, whereas too little M13 primer may result in low fluorescent signal.

Of 19 primer pairs tested using a unique optimized set of PCR conditions in all mammalian samples, including a 1:2 (forward primer:M13-primer) ratio, nine pairs yielded a scorable band pattern, e.g. C2-1218 (Figure 4.8). Failure for 10 of the 19 pairs may be a consequence of the addition of the universal, fluorescently labelled M13 primer, which increases considerably the amount of primer dimers (as seen on Figure 4.8) and may result in limited or no amplification.

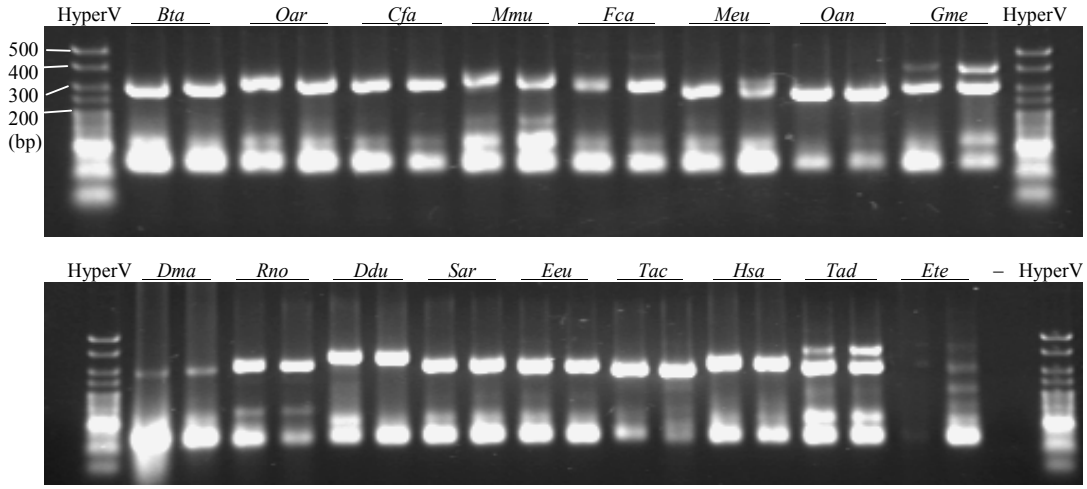


Figure 4.8: PCR amplification results for C2-1218 in 17 mammals and a water-only negative (–) control (with M13 primers). *Bta*: cow, *Oar*: sheep, *Cfa*: dog, *Mmu*: mouse, *Fca*: cat, *Meu*: tammar wallaby, *Oan*: platypus, *Gme*: pilot whale, *Dma*: quoll, *Rno*: rat, *Ddu*: dugong, *Sar*: shrew, *Eeu*: hedgehog, *Tac*: echidna, *Hsa*: human, *Tad*: dolphin, *Ete*: tenrec.

4.4.7 Cross-species genotyping

Nine microsatellites were genotyped at least once for all samples, although results from chimpanzee were unavailable at the time of writing. Table 4.4 shows allele range and number of alleles identified at all loci for all species including the number of individuals successfully genotyped. PCR amplifications and subsequent genotyping were particularly successful at five of the nine loci, namely C2-1218, C2-1915, C4-1514, C9-1918 and C17-4243 (Table 4.3), but we were also able to obtain partial results for the four remaining loci.

Table 4.4: Intraspecies polymorphism (range of allele length) at conserved mammalian microsatellite loci using comparative degenerate primers. Number of alleles and number of diploid individuals successfully genotyped are shown in brackets. Asterisks (*) indicate sequenced loci. NG: no genotype available. n/a: data not available at the time of writing. Locus C6-1112 in dog showed a 3-peak pattern at all individuals successfully genotyped.

	C2-1218*	C2-6868*	C2-1915*	C4-1514*	C6-1112	C9-1918	C14-9692	C15-3531	C17-4243*
Human	268-294 (9/18)	228 (1/20)	166-178 (5/17)	281-283 (2/20)	152-156 (2/19)	300-302 (2/14)	234-237 (3/20)	226-228 (2/17)	311 (1/20)
Chimpanzee	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mouse	291-301 (10/20)	242-291 (16/17)	216-238 (10/14)	313-317 (4/19)	155-161 (3/18)	311 (1/20)	240-242 (2/19)	297-299 (2/16)	319-325 (5/19)
Rat	274-280 (2/20)	236 (1/20)	176-180 (3/20)	274 (1/20)	158-162 (3/19)	NG	NG	237 (1/20)	326 (1/15)
Dog	268-278 (9/20)	256-268 (4/18)	180-191 (5/19)	297-299 (2/17)	3 peaks (1/11)	NG	214 (1/20)	NG	309-310 (2/20)
Cat	265-276 (8/20)	NG	176-188 (6/19)	NG	NG	NG	NG	NG	312 (1/20)
Cow	259-264 (2/18)	231 (1/20)	167-169 (2/20)	281 (1/20)	146 (1/20)	300-305 (2/20)	208 (1/20)	240-242 (2/20)	308 (1/20)
Sheep	270-280 (8/19)	229-237 (4/14)	163-173 (4/15)	291-292 (2/20)	146 (1/20)	307-308 (2/20)	208-212 (3/18)	NG	305-306 (2/20)
Dolphin	264-278 (4/19)	NG	160-176 (7/19)	291-295 (2/16)	148-150 (2/19)	313-319 (3/16)	214-215 (2/19)	226 (1/20)	303-304 (2/20)
Pilot Whale	265 (1/20)	243 (1/20)	161-174 (6/17)	292 (1/19)	148 (1/20)	313-317 (4/18)	216 (1/20)	223 (1/20)	307 (1/20)
Hedgehog	260-272 (5/20)	225-230 (5/20)	168-172 (3/20)	321-325 (2/20)	148-154 (2/20)	345 (1/20)	151-157 (2/20)	213-227 (7/20)	303 (1/20)
Shrew	309-329 (11/20)	254-256 (3/20)	221-223 (2/20)	281 (1/20)	NG	NG	NG	NG	309-313 (4/19)
Dugong	269-273 (4/20)	225 (1/17)	176 (1/20)	274 (1/19)	138 (1/17)	289 (1/18)	NG	222 (1/17)	294-298 (3/18)
Tenrec	NG	NG	NG	281 (1/20)	NG	NG	NG	NG	316-319 (4/15)
Tammar wallaby	249-291 (9/16)	NG	193-195 (2/16)	281 (1/16)	149 (1/16)	NG	NG	191-283 (14/15)	325-332 (5/10)
Quoll	241-243 (2/9)	318-342 (6/8)	NG	295 (1/20)	148 (1/20)	NG	203 (1/14)	NG	299 (1/15)
Platypus	245-263 (2/15)	346-382 (7/13)	214-226 (4/15)	NG	145 (1/11)	NG	208 (1/18)	NG	298 (1/15)
Echidna	248-252 (4/15)	372-376 (4/13)	NG	317 (1/14)	142 (1/12)	278 (1/20)	205-213 (5/14)	194-196 (2/12)	298 (1/17)

By collecting individuals for each species that were, to our best knowledge (except for rat and pilot whale, see below), unrelated, our expectation was to observe intraspecies polymorphism on top of the interspecies polymorphism identified from the sequence alignments. Overall, there was an interspecies and interlocus variability for polymorphism content. Three loci (i.e. C2-6868, C2-1915 and especially C2-1218) showed high polymorphism (Table 4.4), indicating that they may well be suitable for cross-species applications, whereas other loci generally revealed little or no intraspecies polymorphism.

Several reasons can be proposed to explain this difference of variability between loci, including selection, number of degeneracies in primers, primer G+C content, and length of the longest pure repeat tract. First, it was expected that the increasing number of degeneracies in one or both primers in each primer pair would hamper perfect annealing to target sequences, thus amplification success. However, there was no significant difference in amplification success between highly and slightly degenerate primer pairs (Table 4.3), e.g. highly successful C2-1218 primers both contained two degenerate sites, whereas only one primer in the broadly ineffective C14-9692 pair contained a single degenerate site. Likewise, no difference in G+C content of primer sequence was observed between successful and unsuccessful primer pairs. Alternatively, there might be selective influences as either (i) microsatellites comparatively more polymorphic could have been indirectly selected for mutability (King and Kashi 2007), thus promoting rapid changes for plasticity and adaptive advantages, or (ii) conserved microsatellites showing interspecies polymorphism but low or no intraspecies polymorphism may be located in regions selected for stability (Ackermann and Chao 2006). Finally, empirical studies as well as theoretical predictions have shown that long and pure microsatellites are more polymorphic than short and/or degenerated microsatellites (reviewed in Buschiazzo and Gemmell 2006), which might explain observed differences. Indeed, inspection of the microsatellite structure of

genotyped loci suggests that the differential intraspecies polymorphism is likely to be influenced by the length of pure repeat segments within the microsatellite sequence. The polymorphic C2-1218 locus contained long pure tracts of (CA) motifs in most species used for genotyping (Figure 4.3), except in platypus (which had only four repetitions and accordingly exhibited the least polymorphism). The widely polymorphic C2-1915 locus, despite imperfections in the microsatellite sequence, generally contained at least one long pure sub-unit (>8 repeats), e.g. in mouse where the locus showed the most variability (Figure 4.5, Table 4.3). The C2-6868 locus showed less variability than the two previously described loci, and contained many sub-units of short size (<8 repeats). An exception to this pattern was mouse, which showed a highly expanded tract coupled with extensive polymorphism (Figure 4.4). In contrary to these three loci, other conserved microsatellites showed little or no variation, e.g. C4-1514, which is highly degenerated, and C17-4243, which contains small subunits. Altogether, these observations confirmed the propensity of long pure microsatellite tracts to be more polymorphic than short and/or degenerated microsatellites.

Nevertheless, in contrast to this expectation, we observed a few exceptionally long tracts that did not yield variable allele lengths, e.g. in rat at the C17-4243. This discrepancy might be explained by inconsistencies in the quality of sample sets. For example, although unrelated individuals were sought for, little or no information on the exact origin of samples was available for many species. For example, human showed almost no variation in successfully genotyped loci, although the sequences contain potential highly polymorphic microsatellites. As for rats, our samples were collected from inbred insular populations; therefore the observed low polymorphism of otherwise long and pure alleles might not be surprising. Although dugong, pilot whale, quoll and echidna also showed

little polymorphism, no sequence information is available at this stage to verify whether microsatellites contained short and/or degenerated sequences. Noteworthily, most echidna samples originate from road kills in Kangaroo Island (Australia), and as no previous population-based genetics studies have been carried out on these samples, the possibility that closely related individuals were sampled is conceivable (P. Rismiller, pers. comm.). In addition, pilot whale samples were collected from pod strandings (M. Oresmus, pers. comm.), thus likely represent a matrilineal group where little polymorphism may be expected (Amos et al. 1993).

We also found interspecies variation in the genotyping success. Shrew, cat and tenrec samples did not yield any result for four, six and seven loci, respectively. Sequences from these species were available to design comparative primers, and thus provided ground to expect amplification successes similar to that of other species. Consequently, the quality of DNA extracts for shrew, cat and tenrec may be questionable. Because spectrophotometric quantifications did not show significantly higher measures of DNA impurity for these samples, it is possible that the DNA samples were highly degraded.

Overall, of nine primer pairs employed for cross-species genotyping, five were successful in providing allele length data at the population level across most species, which was the main motive for including them for the sequencing implementation (Table 4.3), and three amplified fragments showing significant intraspecies polymorphism (C2-6868, C2-1915 and C2-1218). We believe that these three loci are strong candidates for cross-species analyses across the Mammalia, and may prove invaluable in many areas of research, including molecular ecology and population genetics.

4.4.8 DNA sequencing to explore sources of variation

For most applications where microsatellites are used as genetic markers, sequencing of loci is not classically undertaken. Generally, once allele lengths are scored, the amount of change provided by the microsatellite is usually assumed to be high enough to dissipate the biased effect of any rare indels in the flanking sequences that may create size homoplasy (Estoup et al. 2002). However, sequence information is necessary to infer correctly the evolution of individual microsatellites (Zhu et al. 2000), to assess the effect of size homoplasy on population structure estimates (Angers et al. 2000; Adams et al. 2004), to construct phylogenetic trees based on conserved microsatellite flanking sequences (Domingo-Roura et al. 2005), or to select informative, well-described and consistent genetic markers for routine applications, e.g. in forensic analyses (Butler 2005).

To inspect whether allele length variations were attributable to additions/deletions of motifs in the present set of conserved microsatellites, four homozygous allele variants (where available) were tentatively sequenced for each species at five loci (Table 4.3, Table 4.4). Tables 4.5-9 present an overview of these results, with total fragment length, microsatellite length and microsatellite sequence given for all variants of successfully sequenced individuals. Success in sequencing the microsatellite sequence was unfortunately less frequent that sequencing failure (143 *vs.* 197, respectively, hence a success rate of 42%), regardless of previous genotyping success. In particular only short nucleotide stretches in flanking sequences could be safely read, which resulted in this information being excluded from the present analysis. A number of reasons can be advanced to explain the relatively low sequencing success: (i) degenerate primers may produce weak or no sequencing results as a consequence of the decreased amount of

primers in the mixture that perfectly match target sequences in the sequencing PCR (Murphy and O'Brien 2007), (ii) two purification procedures were necessary to eliminate the high quantity of primer dimers, thus simultaneously reducing the amount of product to sequence, (iii) no cloning of PCR products was carried out to reduce costs and labour.

When sequences were produced (Table 4.5-9), similar changes observed between total fragment length and microsatellite length between individuals of the same species indicated whether allele length change was attributable to microsatellite variability. Of 40 intraspecies comparisons between allele variants, 10 showed a discrepancy between total length and microsatellite length; a discrepancy that likely stems from short indels occurring in flanking sequences. We could not conclude on this hypothesis because flanking sequence information was mostly absent. Nevertheless, 30 comparisons revealed length changes consistent between total fragment and microsatellites; in all cases, addition/removal of one or several motifs occurred in the longest pure tract(s) of dinucleotide repeats. Polymorphism occurring in compound microsatellites and involving long portions of non-dinucleotide motifs was also observed in mouse at the C2-1218 locus (Table 4.5, CCACACC motif derived from a CA motif) and in platypus at the C2-1915 locus (Table 4.7, CCT motif derived from a CT motif). Three cases of homoplasy (identical size, but different sequence) were observed: in cow at the C2-1218 locus (Table 4.5), in rat and in shrew at the C4-1514 locus (Table 4.8).

Overall, sequencing showed that the majority (75%) of allele length variation could be attributed to mutation occurring in the microsatellite sequence; these loci could therefore be potentially employable as polymorphic genetic markers. In addition, they provide an exceptional framework to infer microsatellite evolution above the species level, not only in closely related species, but across the Mammalia.

Table 4.5: Allele and microsatellite length variation at C2-1218 (Inter- and intraspecies)

Species	Indiv	L_{allele}	L_{micro}	Sequence
Human	1	280	48	$(CA)_{24}$
	2,4	+1	0	$(CA)_{24}$
	3	-2	-2	**$(CA)_{23}$**
Chimp	1,3	*n/a*	50	$(CA)_{25}$
	2		-8	**$(CA)_{19}$**
	4		+4	**$(CA)_{27}$**
Mouse	1	291	64	$(CA)_{18}(CCACACC)_3C(CA)_3$
	2	+5	+5	**$(CA)_{17}(CCACACC)_4$**$C(CA)_3$
	3	+9	+9	**$(CA)_{19}(CCACACC)_4$**$C(CA)_3$
	4	+7	+7	$(CA)_{18}$**$(CCACACC)_4$**$C(CA)_3$
Rat	1,2,3,4	276	49	$(CA)_{17}CTA(CA)_6$
Dog	1	274	44	$(CA)_4TA(CA)_{13}CG(CA)_3$
	2,4	-6	-6	$(CA)_4TA$**$(CA)_{10}$**$CG(CA)_3$
	3	-4	-4	$(CA)_4TA$**$(CA)_{11}$**$CG(CA)_3$
Cat	1	265	38	$(CA)_{19}$
	2,3,4	+6	+2	**$(CA)_{20}$**
Cow	1	269	28	$(CA)_{12}(GA)_2$
	3,4	0	0	**$(CA)_{13}GA$**
Sheep	1	276	43	$(CA)_{11}CG(CA)_4(C)_3(CA)_4$
	2,4	-6	-6	**$(CA)_{13}$**$(C)_3(CA)_4$
	3	+2	+2	**$(CA)_{12}$**$CG(CA)_4(C)_3(CA)_4$
Dolphin	1	264	32	$(CA)_{16}$
	2	+4	+4	**$(CA)_{18}$**
Pilot Whale	1,2,3,4	265	34	$(CA)_8TA(CA)_8$
Hedgehog	2	260	40	$(CA)_{20}$
	3	0	+4	**$(CA)_{22}$**
	4	+8	-2	**$(CA)_{19}$**
Dugong	1	270	38	$(CA)_4CG(CA)_{14}$
	2,4	+1	+2	$(CA)_4CG$**$(CA)_{15}$**
	3	+2	0	$(CA)_4CG$**$(CA)_{15}$**
T. wallaby	1,3	251	41	$(CA)_{13}CG(CA)_5ACA$
Platypus	2,3,4	245	12	$(CA)_4(A)_4$

Table 4.6: Allele and microsatellite length variation at C2-6868 (Inter- and intraspecies)

Species	Indiv	L_{allele}	L_{micro}	Sequence
Human	1,2	228	42	$(CT)_5TT(CT)_4(T)_4C(CT)_4CC(CT)_2TTCT$
Chimp	1,2,3,4	*n/a*	42	$(CT)_5TT(CT)_4(T)_4C(CT)_4CC(CT)_2TTCT$
Rat	2,3	236	50	$(CT)_7CC(CT)_6(TC)_2(CT)_4CC(CT)_4$
Dog	4	262	76	$(CT)_8CCTT(CT)_{12}TT(CT)_3TTCC(CT)_5CC(CT)_4$
Cow	1,2,3,4	231	44	$(CT)_5TT(CT)_5(T)_4(CT)_4CC(CT)_4$
Sheep	2	231	44	$(CT)_5TT(CT)_5(T)_4(CT)_4CC(CT)_4$
	4	-2	-2	$(CT)_5TT$**$(CT)_4$**$(T)_4(CT)_4CC(CT)_4$
Dolphin	3,4	*n/a*	56	$(CT)_5GTC(T)_3(CT)_9(T)_3C(CT)_4CC(CT)_4$
Pilot Whale	3	243	56	$(CT)_5GTC(T)_3(CT)_9(T)_3C(CT)_4CC(CT)_4$
Hedgehog	3,4	230	44	$(CT)_5AT(CT)_4(T)_3C(CT)_5CC(CT)_4$
Shrew	1,4	256	68	$(CT)_3TT(CT)_7TT(CT)_4CCTT(CT)_5TCTT(CT)_4CC(CT)_4$
	2	-2	0	$(CT)_3TT(CT)_7TT(CT)_4CCTT(CT)_5TCTT(CT)_4CC(CT)_4$
Dugong	1,3,4	225	40	$(CT)_5TT(CT)_3(T)_3C(CT)_4TC(CT)_4$

Table 4.7: Allele and microsatellite length variation at C2-1915 (Inter- and intraspecies)

Species	Indiv	L_{allele}	L_{micro}	Sequence
Human	1,2,3,4	283	49	CTGTGC$(CT)_{14}$TT$(TC)_5$TTC
Chimp	1,2,3	*n/a*	37	CTGTGC$(CT)_{14}$TTC
	4		+2	CTGTGC**$(CT)_{15}$**TTC
Rat	2,3,4	273	57	$(CT)_2$GC$(CT)_{25}$C
Cow	1,3,4	169	50	$(CT)_2$GC$(CT)_2$T$(CT)_8$CC$(CT)_9$TTC
	2	-2	0	$(CT)_2$GC$(CT)_2$T$(CT)_8$CC$(CT)_9$TTC
Sheep	2,4	181	58	$(CT)_2$GC$(CT)_2$T$(CT)_{12}$CC$(CT)_9$TTC
	3	-8	-8	$(CT)_2$GC$(CT)_2$T**$(CT)_{10}$**CC**$(CT)_8$C**
Dolphin	1,3,4	172	47	$(CT)_{22}$TTC
	2	-4	-4	**$(CT)_{20}$**TTC
Hedgehog	1,4	168	49	$(CT)_2$GCCTGCTT$(CT)_{15}$TTC(T)3C
	2,3	+2	+2	$(CT)_2$GCCTGCTT**$(CT)_{16}$**TTC(T)3C
Shrew	1	280	99	CAT$(TC)_2$TG$(TC)_2$TG$(CCTCT)_2$GC$(CTCTGT)_2(CT)_4(GTCTCT)_2(CTGTCT)_2$\\ \\$(CT)_6(C)_3(T)_3(CCCTCT)_1(CT)_2$TTC
	3	0	+2	CAT$(TC)_2$TG$(TC)_2$TG$(CCTCT)_2$GC$(CTCTGT)_2(CT)_4(GTCTCT)_2(CTGTCT)_2$\\ **$(CT)_7$**$(C)_3(T)_3(CCCTCT)_1(CT)_2$TTC
Dugong	1,2,3,4	274		$(CT)_2$GC$(CT)_8$GC$(CT)_3$TT$(CT)_3$CC$(CT)_7$TTC
Platypus	1,4	223	103	CTGTTC$(T)_6C(A)_5(T)_5$ATC$(CCT)_8(CTT)_2T(CT)_2T(CT)_2CC(CT)_4TT(CT)_5(T)3A(T)_7$CCTC
	2	-10	-9	CTGTTC$(T)_6C(A)_5(T)_5$ATC**$(CCT)_5$**$(CTT)_2T(CT)_2T(CT)_2CC(CT)_4TT(CT)_5(T)3A(T)_7$CCTC
	3	-3	-3	CTGTTC$(T)_6C(A)_5(T)_5$ATC**$(CCT)_7$**$(CTT)_2T(CT)_2T(CT)_2CC(CT)_4TT(CT)_5(T)3A(T)_7$CCTC

Table 4.8: Allele and microsatellite length variation at C4-1514 (Inter- and intraspecies)

Species	Indiv	L_{allele}	L_{micro}	Sequence
Human	1,2,3,4	283	72	GA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_4$CGC$(A)_3(CA)_8$CT$(CA)_2$AACA
Chimp	1,2	*n/a*	70	GA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_4$CGC$(A)_3(CA)_{12}$AACA
	3		+2	GA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_4$CGC$(A)_3$**$(CA)_{11}$**AACA
	4		-1	GA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_4$CGC$(A)_3$**$(CA)_9C(AACA)_2$**
Mouse	1,4	315	104	GATA$(CA)_5$AAGACT$(CA)_2$TA$(C)_4$TGTA$(CA)_2$TACATTCCT(GCAC)$_2$A$(CG)_3(CA)_4$// //$(CT)_2(CA)_{10}$GTCTCG$(CA)_2$
	2,3	+2	+2	GATA$(CA)_5$AAGACT$(CA)_2$TA$(C)_4$TGTA$(CA)_2$TACATTCCT(GCAC)$_2$A**$(CG)_4(CA)_3$**// //**CT$(CA)_{12}$**GTCTCG$(CA)_2$
Rat	1,2,3	274	66	GATA$(CA)_6$AAGACT$(CA)_3(C)_3$TTGTA$(CA)_3$CC$(CA)_5$CC$(CA)_5$
	4	0	0	GATA$(CA)_6$AAGACT**$(CA)_4$C**TTGTA$(CA)_3$CC$(CA)_5$CC$(CA)_5$
Dog	1,2,3,4	297	86	GA$(CA)_3$AA$(CA)_3$CC$(TG)_2(CA)_2$TA$(CA)_8$CGCCAACG$(CA)_2$TA$(CA)_3$CG$(CA)_2$TA// //$(CA)_5$AACA
Cow	1,2,3,4	280	70	GA$(CA)_3(A)_4(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_2$CGC$(A)_3(CA)_3$TACACGCTCACG$(CA)_2$TA// //CA$(C)_3$G
Sheep	1,2,3,4	292	80	TAGA$(CA)_3(A)_4(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_2$CGC$(A)_3(CA)_3$TACACGCTCACGA$(TA)_2$// //$(CA)_3$TACA$(C)_3$G
Dolphin	2	291	79	CAGA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_6$AACACGCATA$(CA)_{10}$TA$(CA)_3$
	3,4	0	0	CAGA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_6$AACACGCATA**$(CA)_7CG(CA)_2$**TA$(CA)_3$
P. Whale	1,2,3	292	79	CAGA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_6$AACACGCATA$(CA)_7$CG$(CA)_2$TA$(CA)_3$
Hedgehog	1,3	322	117	$(CA)_5(A)_3$TA$(CA)_3$CC$(T)_3(GC)_2$AC$(GC)_2$GT$(GC)_3$A$(CA)_2$TA$(CA)_{15}$TATGC$(A)_3(CA)_7$// //TA(CA)3AATA
	2,4	-2	-2	$(CA)_5(A)_3$TA$(CA)_3$CC$(T)_3(GC)_2$AC$(GC)_2$GT$(GC)_3$A$(CA)_2$TA**$(CA)_{14}$**TATGC$(A)_3(CA)_7$// //TA$(CA)_3$AATA
Shrew	1,2,4	281	69	CAGA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_4$CG$(A)_3(CA)_4$CG$(CA)_2$TA$(CA)_3$GACA
	3	0	0	CAGA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$**TACACG$(CA)_2$**CG$(A)_3(CA)_4$CG$(CA)_2$// //TA$(CA)_3$GACA
Dugong	1,2,3,4	274	67	CAGA$(CA)_3$AATA$(CA)_2$CC$(TG)_2(CA)_2$TA$(CA)_3$AA$(CA)_4$CT$(CA)_3(CG)_2$CATAC$(A)_3$CA
Tenrec	1,2,3,4	281	64	AGA$(CA)_4$AA$(CA)_3$CG$(TG)_2(CA)_2$TCCAC$(GC)_2(A)_3$CACG$(CA)_2$CTCACG$(CA)_4$GATG
T. wallaby	1,2,3,4	281	73	C$(A)_3(CA)_5$CATA$)_2$CCT$(TG)_3(CA)_2$CG$(CA)_2$TG$(CA)_2$CT$(CA)_2(CG)_2(CA)_6$TCAG
Quoll	1,3,4	295	89	CAGA$(CA)_3$TGCCT$(G)_3$TG$(CA)_2(C)_5$GCCA$(G)_3$CA$(C)_3$ACGTG$(CA)_2$CG$(CA)_4(C)_3$ACG(CGC A$)_2$CACG$(CA)_4$CC
Echidna	1,2,3,4	317	96	$(CA)_3$GACATACCAACG$(CA)_2$TA$(CA)_4$TGCAGGCACG$(CA)_3$GACACG$(CA)_2$C$(CA)_2(CCCA)$ $_2$CA$(CCCACA)_2(CA)_3$GAACA

Table 4.9: Allele and microsatellite length variation at C17-4243 (Inter- and intraspecies)

Species	Indiv	L_{allele}	L_{micro}	Sequence
Human	1,2,3,4	311	56	$(TTTC)_2(TC)_7CC(TC)_5(C)_4(TC)_3(T)_8(TC)_2$
Dolphin	1,2,3,4	303		$(TC)_2(T)_4(TC)_5CC(TC)_4(C)_4(TC)_3(T)_7(TC)_2$
Tenrec	1,2,4	316	61	$TGTC(T)_4(TC)_5CC(TC)_5(C)_4(TC)_3CC(TC)_4(T)_7(TC)_2$
	3	+2	+2	$TGTC(T)_4\mathbf{(TC)_6}CC(TC)_5(C)_4(TC)_3CC(TC)_4(T)_7(TC)_2$
Echidna	1,2,3,4	298	43	$(TC)_2(T)_4(TC)_4CCTCGC(TC)_2CC(TC)_3(T)_6CTT$
Platypus	1,2,3,4	298	43	$(TC)_2(T)_4(TC)_4CCTTGC(TC)_2CC(TC)_3(T)_6CTT$

4.5 Discussion

Drawing on the recent multiplication of genome sequencing projects, independent but similar methods to help develop, design and implement comparative primers in mammalian species have been published (Housley et al. 2006; Murphy and O'Brien 2007), but none had an emphasis on the use of conserved microsatellite markers for cross-species studies. Here, we used a comparative genomic approach to identify, develop and implement cross-species primers for microsatellites conserved across the Mammalia. Mammals were chosen because (i) a significant number of mammalian genomes have been released for public use (e.g. 12 at the UCSC Genome Browser, 25 at the Ensembl Genome Browser; as of 18/04/2008), including marsupial (opossum) and monotreme (platypus) species; (ii) wide-ranging whole-genome alignments have been produced and are publicly available, such as the UCSC 17- and 28-WA (Miller et al. 2007); (iii) the basal diversification of mammalian species occurred ~160 Myr ago and an explosion of diversification occurred around the K/T (Cretaceous/Tertiary) boundary, ~65 Myr ago (Bininda-Emonds et al. 2007). Evolutionary distance between compared species was therefore much larger than that of species that are typically chosen for comparative primer design (5-10 Myr, Gemmell et al. 1997, although see FitzSimmons et al. 1995 and Rico et al. 1996). Focusing the present analysis on mammals thus ensured a large evolutionary scope as well as a solid framework to identify broadly conserved microsatellites, but also

critically decreased the number of comparative primers to potentially identify and design, as substitutions accumulate over time in microsatellite flanking sequences.

Nevertheless, of ~1000 randomly selected, widely conserved dinucleotide microsatellites, 19 (2%) were suitable to design degenerate comparative primers potentially useful to genotype and sequence fragments <350 bp, which is a remarkable feat considering the breadth of the Mammalia. In addition, our initial random subset represents only a fraction of all microsatellites that were identified across a wide range of mammalian species. For example, using the most comprehensive dataset of conserved mammalian microsatellites (Chapter 2), we were able to find 4084 human dinucleotide repeats conserved in at least five non-primate mammals. By extrapolation, we estimate that at least 80 loci should be suitable for primer design using this selection criterion, and we anticipate that more should be identified under less stringent conditions (e.g. conservation in human-mouse-opossum). Moreover, other types than dinucleotide repeats may also be used for cross-species transfer of microsatellite markers, e.g. tetranucleotide markers, which are also conserved in high numbers in mammalian genomes. Furthermore, if there is success in designing comparative primers across all Mammals, then many more are expected to be developed from comparisons within subgroups of the Mammalia, making this new collection of conserved microsatellites (see Chapter 2) a precious source for future cross-species studies. For example, conserved microsatellites could help examine the under-studied population structure of the long-beaked echidna by bypassing the step of marker development. To this end, comparisons could be limited to microsatellites conserved in platypus, opossum and an outgroup reference species such as human, because platypus share more microsatellites with opossum than with any other mammals studied to date (Chapter 3) and the use of a more distant species would allow the identification of fairly stable primer sites.

In the present study, amplification, genotyping and sequencing success rates were often inconsistent. Of 19 designed comparative primers, nine were successfully optimized and five were suitable for genotyping and sequencing. A number of methodological choices were made to decrease costs, but they may have accentuated failure rates, e.g. Chelex extraction method (impure DNA), M13-genotyping (primer dimers, inconsistent fluorescent signal), use of degenerate primers and no cloning for sequencing PCR (weak or no sequence reads). In addition, we had little or no control on sampling and DNA quality for most of our samples, which may have had detrimental consequences on the overall quality of our results. Drawing on these experiences, guidelines are outlined in Box 4.1 to help others planning to use conserved microsatellites to develop comparative primers.

Overall, our comparative primers still yielded good genotyping results for five of the nine fully optimized loci. Intraspecies polymorphism was strongly associated with length and purity of repeat tracts, which emphasized the importance of examining the sequence structure of microsatellites to select polymorphic genetic markers. Our attempt to use comparative primers to sequence the region of interest in many species was therefore justified, especially in species whose genomes are not sequenced (dolphin, pilot whale, dugong, quoll and echidna). Sequence information demonstrated that most changes (75%) in total fragment length at the five loci were attributable to mutations in the microsatellite sequence rather than in the flanking sequences, suggesting that comparative primers designed for these loci are invaluable candidates for being employed as universal genetic markers across the Mammalia.

Box 4.1: Guidelines to facilitate the identification, design, optimization and implementation of comparative microsatellite primers.

- Preparation of genomic DNA:
  - Use commercial DNA extraction kits to yield high quality genomic DNA. To reduce costs, regenerate extraction columns for several rounds of extraction (Siddappa et al. 2007);
  - Use aliquots to store DNA extracts and limit the amount of freezing-thawing cycles that may rapidly degrade DNA quality;
  - As far as possible, avoid getting DNA from external sources.
- Identification of conserved microsatellites
  - If comparative primers for wide-ranging species are searched, alignments should contain sequences available from all species included in the same lineages;
  - If comparative primers for one particular species are looked for, alignments should only contain sequences available from closely related species. Alternatively, sequences from one additional, more distant species may help designing more robust primers;
  - Average size and pure microsatellites should be preferred as they are more mutable and evolve through mutational dynamics well explained by current models of evolution (Ellegren 2004). The use of very long microsatellite loci raises issues of upper length constraints and homoplasy (Estoup et al. 2002).
- Identification of microsatellites with potentially conserved priming sites:
  - If a limited number of initial conserved microsatellites were selected, visual assessment on the UCSC Genome Browser is a repetitive but effective approach;
  - Alternatively, a more thorough approach may be develop if many more loci need to be reviewed, e.g. writing a script to identify at least 20 bp with no more than 3 dissimilarities in both side of microsatellites.
- Comparative primer design:
  - Designing degenerate primers using PrimaClade can give relatively good transfer results across species (e.g. C2-1218, Table 2), but this strategy limits the amount of perfect annealing to target sequences, and likely reduces the quality of genotyping and sequencing results;
  - Alternatively, non-degenerate primers where weak-bond mismatches are allowed may be preferred (Murphy and O'Brien 2007), although such mismatches also lowers affinity to target sequences in species where mutations occurred.
- PCR optimization:
  - Touchdown PCR profiles facilitate the optimization process when many primer pairs are tested, but tailored standard profiles may be preferred for individual loci to improve the annealing specificity and amplification success.
- Genotyping
  - The use of M13 primers is only advised to reduce costs when assessing microsatellite polymorphism at the population level
  - Fluorescent primers should be purchased and used when the utility of the comparative primer pair has been shown, because they reduce the amount of primer dimer, improve the consistency of amplifications and yield a higher fluorescent signal for detection on a DNA sequencer.
- Sequencing
  - Cloning should be used to improve the consistency of sequencing results and to access sequence information from heterozygous individuals;
  - When possible, primers should be redesigned to perfectly match the target sequence in the species of interest;
  - Sequencing primers should not be synthesized with M13-tails;
  - Alternative purification methods to filter plates may be required to lessen the lost of product of interest, e.g. commercial spin-column kits and touch-prep (Murphy and O'Brien 2007).

4.6 Acknowledgments

We are indebted to Cathy Riemer and her genomic tool Gmaj, which inspired in the first place the methodology behind the identification of conserved microsatellites using whole-genome alignments. Likewise, this work would not have been possible without all the generous donors who provided mammalian samples. Although many more were involved in the overall effort to gather this collection, we express our warm thanks to J. Hickford, L. Moller, M. Oresmus, S. Baker, I. Vargas-Jentzsch, M. Hale, B. Robertson, G. Yannick, G. Hausser, D, Tautz, A. Amanzadeh, F. Shokri, A. Stone, S. Goodman, A. MacMahon, D. Blair, J. Graves, M. Cardoso, C. Whittington, S. Nicol and P. Rismiller. J. Beck's involvement proved essential to locate some of these samples and achieve the technical side of this study. We also thank L-A Pfister for her work on the chimpanzee samples. M. Hale kindly provided access to her PCR cycler and gave useful technical comments. A. Fouquet, A. Marshall, S. Negro, B. Robertson, T. Steeves, M. Will, and J. Wolff also provided helpful comments.

Chapter 5

5 Length and structural changes in conserved mammalian microsatellites

5.1 Abstract

Human microsatellites conserved in 12 mammals from eutherian, metatherian and prototherian groups have been identified in previous investigations (Chapter 2, 3). This comprehensive dataset provides ground for uncharted aspects of microsatellite evolution above the species level. Here, we focus our investigation on structural change at orthologous microsatellites, including complexity, motif replacement and array length variation. First, we found that there was a high rate of changes from a simple form to a compound form in primate lineages compared to the reverse change, representing a 20:1 difference in the human lineage. Compound microsatellites were more likely to arise from substitutions and subsequent expansion of a derived motif at the end of the repeat array. The majority of new compound motifs derived from base substitutions towards C and G, a finding at odds with the AT mutation bias found throughout the mammalian genome. More drastic changes were observed among an unexpectedly large fraction of simple microsatellites where motifs were replaced on the total length of the array, a transition that necessitates the emergence of a new motif, its expansion and the loss of the ancestral motif repeat. We found striking motif-specific differences in the propensity for motif replacement. In particular AC, ATC, CCG and AAAT motifs were very stable compared to the labile AT and AAAAC motifs. We also found genome-specific differences, with species exhibiting elevated rates of change, e.g. rodents, also showing a comparatively increased proportion of motif replacement. Finally, we were able to identify significant differences in the length distribution of orthologous microsatellites, and used empirical cumulative distribution functions (ECDFs) as a novel tool to find length distribution differences within length ranges. Although most comparisons were

statistically insignificant for microsatellites composed of long motifs (3-6 bp), significant differences were found for mono- and dinucleotide repeats, confirming previous reports than mouse microsatellites tend to be longer than human microsatellites, which in turn are on average longer than chimpanzee microsatellites. Overall, our results provide new insights on how microsatellites evolve over a large evolutionary scale, thus helping understand how genomes evolve. Alternatively, these findings enabled the identification of those types of microsatellites that are less prone to drastic structural change and may therefore be more suitable to develop cross-species markers. In particular, $(AC)_n$, $(ATC)_n$ and $(AAAT)_n$ microsatellites stand out as the best candidates for such applications and, whereas $(AC)_n$ microsatellites are already widely used, the use of the latter two types should be considered and developed.

5.2 Introduction

Microsatellites are abundant and ubiquitous sequences, and exhibit an exceptional variability compared to the bulk of DNA sequences in eukaryotic genomes (Ellegren 2004). These properties make them particularly useful to infer relationships between individuals (Gill et al. 1985), populations (Jarne and Lagoda 1996) or breeds (Rout et al. 2008) compared to the more slowly-evolving mitochondrial and regular nuclear markers, which usually do not have a rivalling power of discrimination at this low level of divergence. Unfortunately, the understanding and modelling of microsatellite mutational processes lag far behind the application of these genetic markers (Schlötterer 2004).

Despite considerable efforts to cover some of these complexities (reviewed in Buschiazzo and Gemmell 2006), the mode and tempo of microsatellite evolution is still ill-defined in currently implemented models of evolution (e.g. Stepwise and Generalized Mutation models, i.e. SMM and GSM), including (i) the proportion and range of multi-step mutations (addition/deletion of several motifs at a time), (ii) directional mutation bias towards an increase in length, (iii) how mutation rate is affected by allele length and sequence composition, (iv) constraints on allele size, (v) whether loci persist indefinitely, and (vi) selective influences at the species level. However, this lack of knowledge can be compensated by selecting and exploiting well-defined microsatellites exhibiting mutational patterns in compliance with available models. As stressed by Pardi et al. (2005), there are significant differences between genomic and marker $(AC)_n$ microsatellites. Marker microsatellites are on average longer, less interrupted, and thus have longer uninterrupted segments. In contrast, most genomic $(AC)_n$ microsatellites are short (<10 repeats, Sibly et al. 2003; Pardi et al. 2005) and they contain more interruptions (Pardi et al. 2005). This difference forms the basis of the selection process to isolate informative markers, as long and pure microsatellites generally provide greater polymorphism (Wierdl et al. 1997; Kruglyak et al. 1998; Sibly et al. 2001; Whittaker et al. 2003; Boyer et al. 2008).

Whereas the isolation of robust markers can be achieved reasonably well for intraspecies application, the heterogeneous mutational dynamics of microsatellites and their flanking sequences between taxa (reviewed in Buschiazzo and Gemmell 2006) and the difficulty to transfer microsatellite markers between species (Barbara et al. 2007) complicate the isolation of well-characterized conserved markers for interspecies applications, e.g. phylogenetics, comparative mapping and cross-species

population genetics. *First*, conservation of the primer sites may allow amplification of a scorable fragment despite the loss of a genuine microsatellite sequence in non-focal taxa (Taylor et al. 1999). *Second*, mutations in the genotyping primer sites can result in failure to amplify fragments through PCR, causing the otherwise conserved locus to appear absent (Barbara et al. 2007). *Third*, the flanking sequences adjacent to the repeat motif may experience indel events, causing either fragment sizes to be out of phase with the expected change in repeat length (Karhu et al. 2000; Shao et al. 2005) or size homoplasies (Estoup et al. 1995; van Oppen et al. 2000). *Fourth*, repeat motifs at orthologous loci can change in complexity, from a simple repeat to one that is interrupted or consisting of multiple repeat motifs (Estoup et al. 1995l; Angers and Bernatchez 1997; Colson and Goldstein 1999; Makova et al. 2000; Zhu et al. 2000; Shao et al. 2005; Lopez-Giraldez et al. 2007), or *vice versa* (Garza and Desmarais 2000; Harr et al. 2000), and homologous simple microsatellites may also have altogether different motif units (Riley and Krieger 2004; Riley and Krieger 2005; Riley et al. 2007). *Finally*, taxon-specific features and selective influences may alter microsatellite mutational dynamics at orthologous loci, even among closely related species (Laidlaw et al. 2007), possibly through interspecific variability in the efficiency of the mismatch-repair machinery (Harr et al. 2002; Li 2008). Altogether, these difficulties led to a general cautionary approach regarding the use of orthologous microsatellite loci when evolutionary distance between focal and non-focal species increases, i.e. > 5-10 Myr (Primmer et al. 1996; Gemmell et al. 1997).

However, the comprehensive identification of microsatellites conserved above the species, genus, or even order levels of the Mammalia (Chapters 2-4) provides a timely framework to remedy this lack of knowledge and exploit their great potential in cross-species investigations.

Of the five points listed above, the first one relates to the life expectancy of microsatellites in genomes, a highly variable trait among microsatellites (Chapters 2 and 3). Selecting loci identified in large subsets of the 13 mammalian species included in the analyses will increase the likelihood that the microsatellite is also present in any other mammalian species, especially if this species is closely related to the taxa composing the subset. The second and third points can be circumvented by identifying conserved microsatellites with mutation-purified flanking sequences, and developing robust cross-species primers (Chapter 4). In this analysis, we address the fourth and fifth points by inspecting broad patterns of microsatellite evolution in 12 mammalian genomes and comparing structural attributes of orthologous microsatellites, e.g. change in complexity, motif replacement and length variability. The evolutionary analysis of orthologous microsatellites in distant species will also help document the concept of microsatellite life cycle (Buschiazzo and Gemmell 2006) and better understand general mechanisms of microsatellite and genome evolution.

5.3 Material and Methods

5.3.1 Identification and classification of conserved microsatellites

Using whole-genome alignments, all human microsatellites conserved in 12 mammalian species were recovered (Chapter 2, 3). Table 5.1 lists taxa included in the analysis. Perfect and imperfect microsatellites were initially identified in all species

using SciRoKo 3.1 (Kofler et al. 2007) with fixed penalty parameters (score: 12, mismatch penalty: 4, SSR seed min. length: 3, SSR seed min. repeats: 3, max. mismatches at once: 3), effectively retaining microsatellites with a minimum of 12 bp or three repeats, e.g. $(A)_{12}$ and $(AAAAT)_3$. This scan resulted in the identification of 1,754,773 human microsatellite segments, but 96,369 microsatellite segments lying in segmental duplications (Bailey et al. 2001) and 842,707 microsatellite segments associated with repeats other than low complexity and simple repeats were excluded from further analysis. Following this treatment, microsatellites were classified relative to their structural complexity. If the sequence 25 bp upstream and downstream of a microsatellite interval did not contain another microsatellite, the microsatellite was classified as 'simple'. Microsatellites were merged and classified as 'compound' if they were 5 bp or less apart from each other or were overlapping by 5bp or less, 'linked' if they were separated by 5 to 25 bp, or 'mixed' if they contained both linked and compound portions. Conservation was asserted when microsatellite positions overlapped in alignments. In the present study, only simple and compound orthologous microsatellites were considered.

5.3.2 Change in complexity

Lineage-specific events of structural change in microsatellite complexity, i.e. simple to compound and compound to simple microsatellites, were identified in human, chimpanzee and rhesus. Figure 5.1 shows the strategy employed to find unambiguous cases of change from simple to compound microsatellites, whereas a reverse strategy allowed the identification of changes from compound to simple microsatellites. This

approach ensured that the change identified was truly in the order: ancestral structure → derived structure.

5.3.3 Motif replacement

Motifs of all human simple microsatellites conserved exclusively with simple microsatellites of 12 mammalian genomes were compared to motifs in orthologous microsatellites. Proportions were calculated for all possible types of motif replacement (or retention).

5.3.4 Variation in length

Subsets of human simple microsatellites conserved in (i) all 12 species, (ii) the Boreoeutheria, (iii), chimpanzee and mouse, (iv) chimpanzee and (v) mouse were constructed to compare allele length distribution between species using non-parametric approaches. Empirical cumulative distribution functions (ECDFs) were obtained for all comparisons using MATLAB (The MathWorks, Inc.). Confidence intervals were calculated using the formula:

$$\epsilon = \sqrt{\frac{1}{2 \cdot n} \times \log\left(\frac{2}{\alpha}\right)}$$

with ε representing the predicted errors above and below the distribution function, n the number of orthologous microsatellites in each dataset, and α set at 0.05.

5.3.5 Statistical analyses

Statistical analyses were performed with the R package (www.r-project.org).

1	Compound	Simple	Simple	Simple
2	Simple	Compound	Simple	Simple
3	Simple	Simple	Compound	Simple

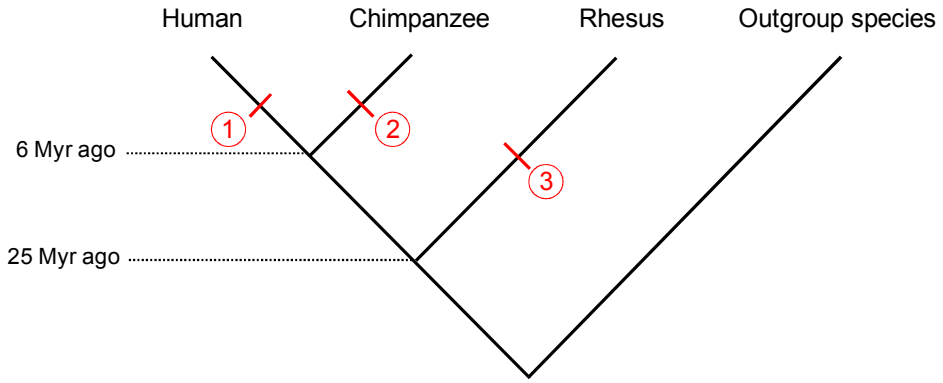


Figure 5.1: Identification of structural changes in primate microsatellites from simple to compound structures. Unambiguous cases of changes in a given primate lineage (1: human, 2: chimpanzee and 3: rhesus) were observed when microsatellites in both other primates and at least one of the outgroup species had a simple structure, but no compound structure in any species. Outgroup species included: mouse, rat, rabbit, dog and cow.

5.4 Results

A broad range of methods have been devised to infer microsatellite evolution (Vargas-Jentzsch et al. 2008). Direct observations of microsatellite mutations (e.g. pedigree analyses and sperm typing) provide a detailed picture of the short-term evolution of hypervariable microsatellites, but lack evolutionary scope and do not provide information for the vast majority of genomic microsatellites (Pardi et al. 2005). Whole genome surveys have provided a means to test on a large scale theoretical models against either microsatellite distribution at equilibrium (Dieringer

and Schlötterer 2003, Calabrese and Durrett 2003) or microsatellite length variability between pairs of species (Sainudiin et al. 2004). Ideally, this can be combined with intraspecies polymorphism data provided by re-sequencing (Brandström and Ellegren 2008), but this opportunity is still exceptional, especially when studying microsatellite evolution in as many as 13 species. To capture unbiased processes of microsatellite evolution in the Mammalia, we sought to compare structural features using subsets of orthologous microsatellites. This strategy allowed us to shed some light on the influences that motif composition and species-specific factors may have on microsatellite evolution.

5.4.1 Identification of orthologous simple and compound microsatellites

Using the dataset of microsatellites conserved between human and 12 other mammals (Chapters 2 and 3), the fraction of exclusively simple orthologous microsatellites was extracted to inspect differences in motif usage and array length between taxa. This resulted in 463,630 human simple microsatellites conserved in at least one species (78% of all human conserved microsatellites). Table 5.1 details the number of simple microsatellites that are orthologous to human simple microsatellites for all species by motif size class.

Table 5.1: Conservation of exclusively simple microsatellites between human and 12 mammalian species.

Species	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-	Total	%All simple	%All conserved
	-nucleotide repeats								
Human	92,252	120,247	76,541	134,881	32,588	7,121	463,630	74.7%	78.0%
Chimp	75,119	106,431	68,471	122,162	29,469	6,309	407,961	73.8%	78.2%
Rhesus	46,016	66,480	40,301	71,987	15,391	3,580	243,755	47.6%	73.2%
Mouse	2,902	8,075	4,146	6,698	1,277	411	23,509	7.3%	55,1%
Rat	1,790	7,605	3,473	5,682	935	305	19,790	8.1%	53.5%
Rabbit	4,530	9,155	4,456	6,574	953	341	26,009	11.3%	64.3%
Dog	10,161	13,974	8,070	12,823	2,457	711	48,196	10.7%	65.1%
Cow	7,955	11,981	6,539	10,050	1,205	289	38,019	11.2%	65.9%
Armadillo	3,755	5,474	4,350	7,055	1,181	301	22,116	11.2%	67.1%
Elephant	8,666	8,208	4,971	7,623	1,052	182	30,702	9.5%	67.3%
Tenrec	1,157	4,697	3,117	4,161	638	177	13,947	8.3%	62.8%
Opossum	758	1,367	1,826	1,613	330	124	6,018	6.0%	59.3%
Platypus	847	842	1,241	676	137	65	3,808	15.3%*	58.6%*

These numbers indicate that, among those human simple microsatellites conserved across the Mammalia, large differences exist in the abundance of each motif class, generally in the order di->tetra->mono->tri->penta->hexanucleotide repeats. Exceptionally, however, tetranucleotide microsatellites were more frequently conserved than dinucleotide repeats in some species, i.e. in primates and armadillo, or less frequently than mononucleotide strings, i.e. in platypus.

Compound microsatellites with orthology in mammalian genomes were also identified. We found 32,067 compound human microsatellites conserved in at least one mammalian species (5.3% of all conserved human microsatellites), a number that compares relatively well with the estimated 10% fraction of compound microsatellites in the human genome (Weber 1990), given that we excluded paralogous microsatellites and microsatellites associated with transposable elements (i.e. 53.51% of our initial dataset of microsatellite segments in human sequences, see Methods). This concordance between genomic and conserved fractions of compound

*Values resulting from integration of platypus microsatellites into the 17-WA framework were used instead of the 6-WA analysis (see Chapter 3).

microsatellites confirms that the retention of microsatellites is mostly a random, hence neutral, process (Chapter 2).

The classification of identified compound microsatellites was not trivial, as their nature can reach various degrees of complexity. First, compound microsatellites may consist of two or more adjacent segments. Most compound microsatellites in our dataset comprised two segments (87.7 % in human), but compound microsatellites with up to 10 adjacent motifs were also found. Second, adjacent segments may not necessarily have motifs of similar length. Although the majority of compound segments had motifs of equal length (71.5% of human two-segment compound microsatellites), motif length was also found to differ between segments (28.5%). Third, a large proportion (32% in human) of microsatellites identified as compound were of identical motif, e.g. A-A, which, in theory, would not meet the definition of a compound microsatellite (Chapter 1, Table 1). Because of our utilization of standardized motifs when searching for microsatellites in DNA sequences, e.g. T = A (Kofler et al. 2007), we could not differentiate between true compound microsatellites, e.g. A-T couples, and simple microsatellites interrupted by a short unique sequence but still classified as compound by our classification procedure. For example, the sequence $A_{12}(CGTTG)A_{12}$ is classified as a compound microsatellite of the form A-A. Therefore, for simplicity and to ensure rigorous comparisons, only two-segment compound microsatellites with different standardized motifs were considered for further analysis, unless stated.

Table 5.2 shows the count of these two-segment microsatellites in the Boreoeutheria. In the three primates, the most common couples were identical, with a general over-abundance of purine-rich (A and G) motifs relative to pyrimidine-rich (C and T) motifs. Following trends observed for simple microsatellites, abundance of

two-segment compound microsatellites conserved in primates followed the order di->tetra->tri->penta->hexanucleotide repeats. A-C compound repeats were only found rarely. Beyond primates, although some associated motifs were still common within size classes (e.g. AC-AG, AAGG-AGGG and A-AAAG), patterns of motif preference were somewhat different to primate sequences, with comparatively more GC-enriched repeats. In particular, non-primate compound microsatellites consisting of an $(AC)_n$ segment and a GC-rich segment were relatively more common than within primates (e.g. AC-CG and AC-ACGC, Table 5.2). Also noteworthy, motif preference within the trinucleotide size class differed largely between primates and other mammals, with more GC-rich compound microsatellites present in the latter group. This dichotomy between primates and taxa more distantly related to human is reminiscent of the overall distinctive composition of microsatellites relative to the extent of conservation in vertebrates (Chapter 2), i.e. widely conserved microsatellites are GC-enriched compared to microsatellites conserved in closely related primates. It is therefore important to keep in mind that numbers presented in Table 5.2 are not representative of the genome content in compound microsatellites, rather the subset of conserved loci.

Table 5.2: Conserved mammalian microsatellites consisting of two segments with different motifs.

		Human	Chimp	Rhesus	Mouse	Rat	Rabbit	Dog	Cow
Total		17,655	14,959	11,973	2,083	1,756	1,456	2,681	1,407
1x-motifs	A-C	14	17	12	3	2	2	8	1
2x-motifs		7,339	6,116	4,974	903	850	734	916	594
	AC-AG	3,184	2,713	2,533	714	673	616	704	379
	AC-AT	2,009	2,500	1,836	101	62	66	131	149
	AT-AG	383	688	358	20	13	25	31	17
	AC-CG	291	214	245	66	101	27	49	48
	AG-CG	1	1	2	2	1	0	1	1
3x-motifs		811	681	596	143	124	87	203	72
	AAG-AGG	167	146	115	17	11	5	33	4
	AAT-ATC	98	80	74	6	8	4	13	10
	ACC-ATC	66	53	51	19	16	13	25	7
	AGC-CCG	32	32	23	18	17	21	24	6
	AAC-AGC	32	19	33	13	7	3	2	14
4x-motifs		2,913	2,422	1,938	251	220	122	319	148
	AAGG-AGGG	890	696	536	21	17	26	41	30
	AAGG-AAAG	329	269	185	21	23	2	37	9
5x-motifs		189	150	69	9	2	3	7	0
	AAGGG-AGGGG	27	26	9	3	0	0	1	0
6x-motifs		28	10	3	1	0	0	2	0
Mixed-motifs		6,361	5,563	4,381	773	560	508	1,226	592
	A-AAAG	473	444	394	65	34	54	153	52
	A-AG	360	334	339	16	7	28	88	57
	A-AT	342	298	157	5	2	4	55	4
	AC-ATAC	250	232	157	30	26	23	33	24
	A-AAG	204	209	227	25	10	13	88	20
	AG-AGGG	193	163	127	63	29	36	39	11
	AC-ACGC	104	80	94	28	67	13	23	25
	AC-ACAG	101	88	97	62	55	13	13	28

5.4.2 Change in complexity

Although ~10% of all human microsatellites have a compound structure, little effort has been undertaken to understand the mechanisms responsible for the genesis of compound repeats in genomes. Compound microsatellites may arise (i) by chance, i.e. if a microsatellite emerges randomly in the immediate vicinity of a pre-existing microsatellite, (ii) through gene conversion (Jakupciak and Wells 2000), which may generate compound microsatellites consisting of two complementary motifs (e.g. CTG-CAG), or (iii) through interruptions and motif alteration at the end of, or within the original repeat array, and subsequent expansion of the new repeat through replication slippage and/or indel-like events (Dieringer and Schlötterer 2003). It is unlikely that random events only explain the relatively high fraction of compound microsatellites, and the great majority of compound microsatellites do not consist of complementary motifs (Table 5.2), therefore (i) and (ii) are probably insignificant processes in the genesis of compound microsatellites. Rather, because point mutation rates are higher in microsatellite sequences than in the rest of the human genome (Sibly et al. 2003; Pumpernik et al. 2008) and imperfections can be duplicated within the array (Harr et al. 2000; Rolfsmeier et al. 2000), the likely molecular mechanisms at the origin of most compound microsatellites is point mutation and further expansion of a derived motif within or at the end of an ancestral motif. In addition, this mutational scheme may help determine how the various structures observed may arise, i.e. single-step slippage/indel-like expansions will generate compound microsatellites in which the ancestral and derived segments have the same motif length (e.g. AC-AG), whereas the motif length of the ancestral and derived segments will differ in multi-step slippage/indel-like expansion events (e.g. A-AT, AC-ATAC and A-AAAG). Our finding that the majority (71.5%) of two-segment compound microsatellites were composed of motifs of equal

length supports this view, as most studies of microsatellite mutability in human have found that single-step mutations predominate (Weber and Wong 1993; Amos et al. 1996; Brinkmann et al. 1998; Kayser et al. 2000; Leopoldino and Pena 2002; Gusmão et al. 2005; Nikitina et al. 2005), but see Huang et al. 2002.

It may also be asked whether derived segments emerge preferentially within or at the end of ancestral segments. Given the large fraction of two-segment compound microsatellites in the form m1-m2 (n = 17,655 in human) compared to three-segment compound microsatellites in the form m1-m2-m1 (n = 1,645), our results imply that derived segments arise significantly more frequently at the end of a pre-existing microsatellite, rather than within its repeat array (χ^2 test from 1:1 expectation, $\chi^2 = 13280.83$, p-value < 0.0001). This polarity in the emergence of compound microsatellites suggests that there may be an associated polarity of point mutations in the repeat array. Although this view is contradicted by an analysis of microsatellite interruptions in human that argued that fewer interruptions occur at the ends of an array than in internal positions (Sibly et al. 2003), it is nevertheless supported by studies in artiodactyls (Brohede and Ellegren 1999) and chicken (Brandström and Ellegren 2008). This contention certainly requires further analysis, but the existence of a form of purification occurring within the array, under which interruptions are eliminated through replication slippage (Harr et al. 2000), helps speculate in favour of polarity of point mutations, thus emergence of compound segments, in microsatellite sequences.

We identified unambiguous cases of emergence of two-segment compound microsatellites from simple microsatellites in each one of the three primate lineages (Figure 5.1). Overall, more compound microsatellites emerged in the lineage leading to rhesus macaque than in

the human and chimpanzee lineages (Table 5.3). Human and chimpanzee lineages diverged ~6 Myr ago, whereas rhesus macaque shared a common ancestor with hominoids ~25 Myr ago (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), corroborating the expectation that more simple→compound structure changes occur with increasing evolutionary time. More changes occurred in the human lineage than in the chimpanzee lineage, but we could not determine whether this finding reflects a genuine difference in the mutational processes acting on human and chimpanzee simple microsatellites, or a consequence of the bias introduced by using the human genome as a reference to find microsatellites conserved in related genomes, i.e. the initial dataset contains more human microsatellites than chimpanzee microsatellites.

The most frequent types of emerged compound microsatellites found in this study generally correspond to the most abundant types present in primate sequences (Table 5.3), e.g. (AC-AG) and (A-AAAG), but we also found significantly higher numbers of certain types of compound microsatellites than types that were more commonly found at the genome scale. In particular, there were significantly more genomic AC-AT microsatellites than genomic AC-CG microsatellites (Table 5.2, e.g. in human: χ^2 test from1:1 expectation, $\chi^2 = 265.87$, p-value < 0.0001), but the observation was reversed for *de novo* compound microsatellites, at least in the larger human and rhesus datasets (Table 5.3, e.g. in human: χ^2 test from 1:1 expectation, $\chi^2 = 13.70$, p-value = 0.0002). This finding, together with the high abundance of AC-AG compound microsatellites (Table 5.3), suggests that GC-enriched compound microsatellites may have been promoted in the recent evolution of microsatellites in primates compared to AT-enriched compound microsatellites, a somewhat unexpected result in view of the general predominance of G:C→A:T *vs*. A:T→G:C changes in mammals (Lipatov et al. 2006).

Table 5.3: Structural change at primate orthologous microsatellites. Simple → Compound changes represent unambiguous cases of the emergence of a compound microsatellite from a simple microsatellite (see Figure 5.1) in one of the primate lineage, whereas Compound → Simple changes represent the loss of one of the segment in a compound microsatellite, resulting in a microsatellite of a simple form.

	Simple → Compound			Compound → Simple		
	Human	Chimp	Rhesus	Human	Chimp	Rhesus
Total	472	149	1,558	23	17	97
A ↔ Compound	42	11	183	-	-	4
A-AAAG	20	1	60			1
A-AG	5	2	27			1
AC ↔ Compound	233	114	608	12	4	45
AC-AG	72	35	272	9	4	33
AC-CG	53	17	85	1	-	1
AC-AT	32	19	57	1	-	8
AC-ACGC	16	10	28	-	-	-
AG ↔ Compound	42	4	125	6	7	13
AC-AG	17	1	53	2	6	9
A-AG	7	-	19	2	-	-
AT ↔ Compound	24	4	111	2	3	8
AT-AC	14	1	76	1	2	5
A-AT	6	1	12	-	-	-
AAG ↔ Compound	6	2	25	2	-	-
A-AAG	5	2	15	-	-	-
AAAG ↔ Compound	12	2	42	1	-	1
A-AAAG	10	1	17	-	-	-

To determine whether certain motifs were more prone to generate a derived segment than others, we compared for each species motif-specific ratios of simple microsatellites giving rise to compound microsatellites to the occurrence in all conserved simple microsatellites (Table 5.4). Only human and rhesus mono- (A) and dinucleotide motifs (AC, AG, AT) were considered, as larger motifs and simple microsatellites in the chimpanzee lineage only gave rise to a restricted number of compound microsatellites (Table 5.3), and were thus not suitable for statistical analysis. If a significant difference exists between these ratios then our interpretation is that one of the motifs is preferentially compounded compared to the other one. We found that, in human, certain motifs were more prone to form compound microsatellites than others, in the order AC>AG=AT>A. In contrast, in rhesus macaque, the order was less significant, i.e. AC=AG=AT>A. The data strongly favour AC becoming interrupted, with the others ocurring at much lower frequencies. This

pattern corresponds in human to the expectation that mutations through deamination of C/G bases occur more often than mutations of A/T bases (Coulondre et al. 1978).

Table 5.4: Motif preference in the creation of compound primate microsatellites. Pairwise comparison of numbers of conserved microsatellites in human or rhesus (N) and simple microsatellites (ancestral state) that became compound microsatellites (derived state in the human or rhesus lineages) in the context of motif composition (n). Pearson's χ^2 value is shown and P-values are coded as follows: 0<***<0.001<**<0.01<*<0.05<not significant (n.s).

	Human				Rhesus			
	A N=94,558 n=42	AC N=82,561 n=233	AG N=47,520 n=42	AT N=28,471 n=24	A N=47,520 n=183	AC N=52,862 n=608	AG N=11,781 n=125	AT N=9,545 n=111
A		AC>A 160.26 ***	AG>A 10.33 **	AT>A 6.48 *		AC>A 3642.49 ***	AG>A 1763.16 ***	AT>A 1808.52 ***
AC			AC>AG 53.52 ***	AC>AT 35.78 ***			AC>AG 0.67 *	AC=AT 0.01 n.s.
AG				AG=AT 0.03 n.s.				AT=AG 0.49 n.s.

In addition, we sought to identify unambiguous structural changes from compound to simple repeats, which would require one of the segments to contract through replication slippage, accumulation of interruptions in this particular segment and/or large deletions in the repeat array. Only few such events occurred in the primate lineage (Table 5.3), corroborating the view that deaths of microsatellites are less frequent than genesis of microsatellites in mammalian genomes (Buschiazzo and Gemmell 2006). Accounting for a generation time of 20 years for human, one compound microsatellite emerged every ~636 generations in the human lineage, whereas one simple microsatellite derived from a compound microsatellite arose every ~13,043 generations.

This relatively high compound microsatellite 'birth rate' raises the question as to whether there may be consequences for the transferability of consistent microsatellite markers in cross-species analyses. However, given the large number of simple microsatellites in genomes (e.g. 201,886 human simple microsatellites conserved in both

primates), the odds of amplifying a compound microsatellite when a simple microsatellite is expected are minimal (up to 0.002 %).

5.4.3 Motif replacement

We have seen that not all orthologous simple microsatellites maintained their simple structure intact after species divergence. A fair question to ask is whether orthologous microsatellites that retained their simple structure also retain their internal structure, i.e. repeat motif composition. Obviously, motif replacement suggests that the new motif emerged and expanded to meet the definition of a microsatellite while the ancestral motif disappeared through deletions, a multi-step process that one may assume occurs rarely. Such events, if generalized, may have consequences when transferring microsatellite markers among species, as microsatellites with different motifs show different mutational dynamics (reviewed in Buschiazzo and Gemmell 2006). In addition, motif replacement may also indicate change of function, assuming that at least a fraction of microsatellites conserved above the species level have functions in genomes.

Microsatellite motif replacements have been previously invoked to explain microsatellite genesis from middle and end poly-A strings of *Alu* repeats (Batzer and Deininger 2002). In addition, Riley and co-authors have published several reports of motif replacement in microsatellites located in eukaryotic UTRs (Riley and Krieger 2004; Riley and Krieger 2005; Riley et al. 2007). However, there is to date no comprehensive examination of motif usage among wide-ranging orthologous microsatellites. To palliate this lack of knowledge, the amount and nature of motif replacement between the human genome and other mammalian genomes was quantified for different motifs of varying size.

To simplify the analysis, only standardised motifs were considered (Kofler et al. 2007). Consequently, cases where a given motif (e.g. GATA) was replaced on the same strand by (i) a similar motif differing by its register (e.g. ATAG) or (ii), its complementary motif (e.g. CTAT) or (iii) a combination of both (e.g. TATC), were not reported. Whereas the first case is not detrimental for the accurate interpretation of results (although frameshift would affect transcription of exonic microsatellites), the latter cases create truly false negative motif replacement. However, as discussed previously, it can be safely assumed that new motifs are directly derived from ancestral motifs through point mutations and subsequent expansion; it is thus fairly improbable that complementary motifs are created only by chance, especially for longer motifs, suggesting that such events may occur at very low frequency. Therefore, the utilization of standardised motifs should not have significant consequences on the interpretation drawn from this analysis.

Also noteworthy, only microsatellites that have an orthology to loci in the human genome were considered for this analysis, rather than microsatellites conserved in all 13 species, which were exceptional occurrences (i.e. 65 ubiquitous microsatellites were identified, two of which showing motif replacement). As a consequence, it was not possible to infer accurately the ancestral motif and the approximate timing of motif replacement for all microsatellites considered. Overall, examination of motif evolution among 463,630 human orthologous microsatellites examined across the mammalian phylogeny nevertheless allowed us to distil valuable and unprecedented information about patterns of motif usage and evolution of conserved simple microsatellites in mammals.

Figure 5.2 depicts the proportions of retained and replaced motifs, superimposed on the current mammalian phylogeny (Miller et al. 2007), for seven types of motifs, including one mononucleotide (A), two dinucleotides (AC and AT), two trinucleotides (AAC and ATC),

one tetranucleotide (AAAT) and one pentanucleotide (AAAAC). These motifs were chosen for display as they either were the most common of their type across the genomes examined (A, AC, AAAT and AAAAC) or showed distinctive patterns of motif replacement (AT, AAC and ATC). Due to low numbers, no hexanucleotide repeat was displayed, but the most common motif of this type (AAAAAC) showed very similar patterns to AAAAC (see below).

A number of general trends can be singled out to grasp common patterns among motifs shown in Figure 5.2. First, as expected from the close relationship among primates, chimpanzee and rhesus macaque orthologues consistently showed similar motif usage to human. Second, despite being the next species closest to human in the present phylogeny, mouse and rat showed the highest rate of motif replacement, confirming the rapid evolution of rodents at the genome scale (Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005). Third, irrespective of motif composition, the most common motif replacements involved a single nucleotide difference between ancestral and derived (standardized) motif, confirming that, in most cases, the most parsimonious explanation is certainly valid to explain motif replacement among orthologous microsatellites.

Human simple mononucleotide repeats were almost exclusively (~99%) represented by poly-A strings. Two distinct patterns were observed at orthologous positions of human poly-A strings: whereas average replacement occurred in most mammals (20-35% of orthologous loci), a significantly larger proportion (55-67%) of mouse, rat, tenrec and opossum orthologues presented different motifs (Figure 5.2). In all cases, substitute motifs were A-rich, indicating that a single substitution in the ancestral microsatellite was at the origin of the derived motif. Two scenarios are conceivable, depending on whether the poly-A string is ancestral or derived, including (i) a new motif created from the poly-A string, e.g. microsatellites associated with *Alu* repeats (Batzer and

Deininger 2002), and (ii) a substitution towards A in an A-rich array. Interestingly, the most common motif replacements almost consistently involved A↔C/G substitutions rather than A↔T substitutions, contradicting again the general AT mutation bias (Lipatov et al. 2006). The high proportion of (A)-motif replacement in opossum may be attributed to the large evolutionary time separating marsupials from human and hence the large number of point mutation events, but platypus orthologues showed average motif replacement, which suggests that evolutionary divergence alone cannot readily explain the motif replacement pattern observed in opossum. As for mouse, rat and tenrec, these species have the fastest evolutionary rate of all 13 species, as illustrated by large branch lengths on an independent mammalian phylogenetic tree based on substitution events (Figure 5.2), confirming that more motif replacements occur with increasing substitution rate.

Among dinucleotide repeats, $(AC)_n$ microsatellites were the most abundant, a property that makes them relatively easy to isolate as *de novo* genetic markers. Strikingly, almost no motif replacement occurred at orthologous $(AC)_n$ microsatellites, suggesting that this motif is particularly resistant, i.e. either few point mutations occur in the array and/or point mutations do not alter the main (AC)-repeat usage. In fact, among human dinucleotide microsatellites, $(AC)_n$ repeats of equal total array length showed fewer imperfections than $(AT)_n$ and $(AG)_n$ microsatellites (Kolmogorov-Smirnov test, $D^- = 0.0118$, $P = 0.99$ and $D^- = 0.1412$, $P = 0.18$, respectively). $(AT)_n$ microsatellites showed a remarkably distinctive pattern relative to $(AC)_n$ microsatellites: in all non-primate eutherian species, a significant proportion (18-40%) of orthologous microsatellites showed an AT↔AC replacement, although the proportion was slightly lower for armadillo (12%). Despite a relatively low evolutionary distance from human (~25 Myr), rhesus orthologues also showed a relatively high proportion (13%) of similar replacements, emphasizing that the rate of this process is fairly elevated. Comparatively, fewer replacements occurred in

opossum and platypus orthologous microsatellites. Combining this observation with the previously demonstrated stability of (AC) motifs implied that the (AT) motif was likely the ancestral motif, and that a generalised AT→AC replacement occurred in eutherian microsatellites. The reason for this unilateral instability is unclear, but might explain why AC-AT are the most common type of compound microsatellites in the human genome (Table 5.2).

Among trinucleotide repeats, $(AAC)_n$ microsatellites were the most common after $(AAT)_n$ microsatellites, but significantly differed from other trinucleotide repeats in terms of their genomic location, with ~99% lying in non-exonic regions compared to 96% for $(ATC)_n$ microsatellites and 55% for $(CCG)_n$ microsatellites. Given that no selective pressure therefore acts on most $(AAC)_n$ as it would on coding trinucleotide repeats to restrain motif length variation, and maintain the codon frame and the amino acid chain structure, it might not be surprising that the great majority of AAC replacements did not involve any other trinucleotide motif. Instead, a large fraction of $(AAC)_n$ microsatellites involved replacement events with (A_nC) motifs or poly-A strings. Strikingly, more elephant orthologues were composed of a poly-A string rather than an AAC motif! Replacements involving motifs of both larger and smaller size suggest that the AAC motif may be an intermediary stage, implying that similar fractions of $(AAC)_n$ microsatellites in a given genome are either ancestral or derived from another microsatellite type. Similar to $(AAC)_n$ microsatellites, $(ATC)_n$ trinucleotide repeats were found in great numbers in non-coding regions of the human genome. Nevertheless, they appeared relatively stable regarding motif composition, suggesting that homogeneous distribution does not necessarily entail no selective pressure. Interestingly, a similar motif stability was observed in other trinucleotide repeats frequently found in exons, e.g. $(CCG)_n$ microsatellites, but this is

likely to reflect a direct consequence of increased selective pressure to maintain motif usage. All major replacements of $(CCG)_n$ microsatellites involved other trinucleotide motifs, lending weight to the expectation that selective influences constrain the motif usage of coding repeats.

The AAAT motif stood out as the most abundant motif forming tetranucleotide microsatellites; this can be mostly explained by the large numbers of $(AAAT)_{<4}$ microsatellites present in mammalian genomes, i.e. 52.4 % of all $(AAAT)_n$ microsatellites in our dataset. Despite its high A content, the AAAT motif was very stable in mammalian orthologues (<40% of motif replacement), with motif stability extending through to opossum and platypus; this result contrasts sharply with other A-rich tetranucleotide motifs (i.e. AAAC and AAAG, data not shown), which were significantly more prone to derivation from, or replacement to motifs of different length but similar composition (e.g. AC↔AAAC↔AAC). Interestingly, Kelkar et al. (2008) estimated a significantly lower mutability at equal length for $(AAAT)_n$ microsatellites relative to $(AAAC)_n$ and $(AAAG)_n$ microsatellites. The comparatively higher stability of $(AAAT)_n$ microsatellites, in combination with their large numbers in mammalian genomes, suggests some selective influence that promotes their genesis and prevents their substitutions by microsatellites of different nature and, possibly, functionality. Similar to $(AC)_n$ microsatellites, most conserved $(AAAT)_n$ microsatellites might be ancestral, i.e. they emerged from unique sequence independently from other microsatellite sequence. This view is also supported by the reversed AT mutation bias found in conserved microsatellites compared to the genomewide observations (Lipatov et al. 2006).

Human pentanucleotide and hexanucleotide microsatellites and their mammalian orthologues were poorly represented compared to most microsatellites with smaller repeat units, but $(AAAAC)_n$ repeats were present in adequate numbers to control for significant motif replacements. As expected by its A-rich nature, the AAAAC motif exhibited low stability in mammalian orthologues and was characteristically involved in motif replacements with shorter related motifs (i.e. A, AC, AAC and AAAC).

Overall, this analysis of motif replacement among orthologous mammalian microsatellites demonstrated that not only do microsatellites with different motif length show very distinctive mutational dynamics, but also that the mutational dynamics vary among microsatellites with different motif composition. These results therefore support the earlier studies of microsatellite mutability among motifs of similar size but differing composition (Kelkar et al. 2008, and references therein). The patterns of motif replacement observed here also lend weight to the hypothesis that motif-specific selective forces might affect the stability, thus life expectancy, of microsatellites in related genomes (Buschiazzo and Gemmell 2006).

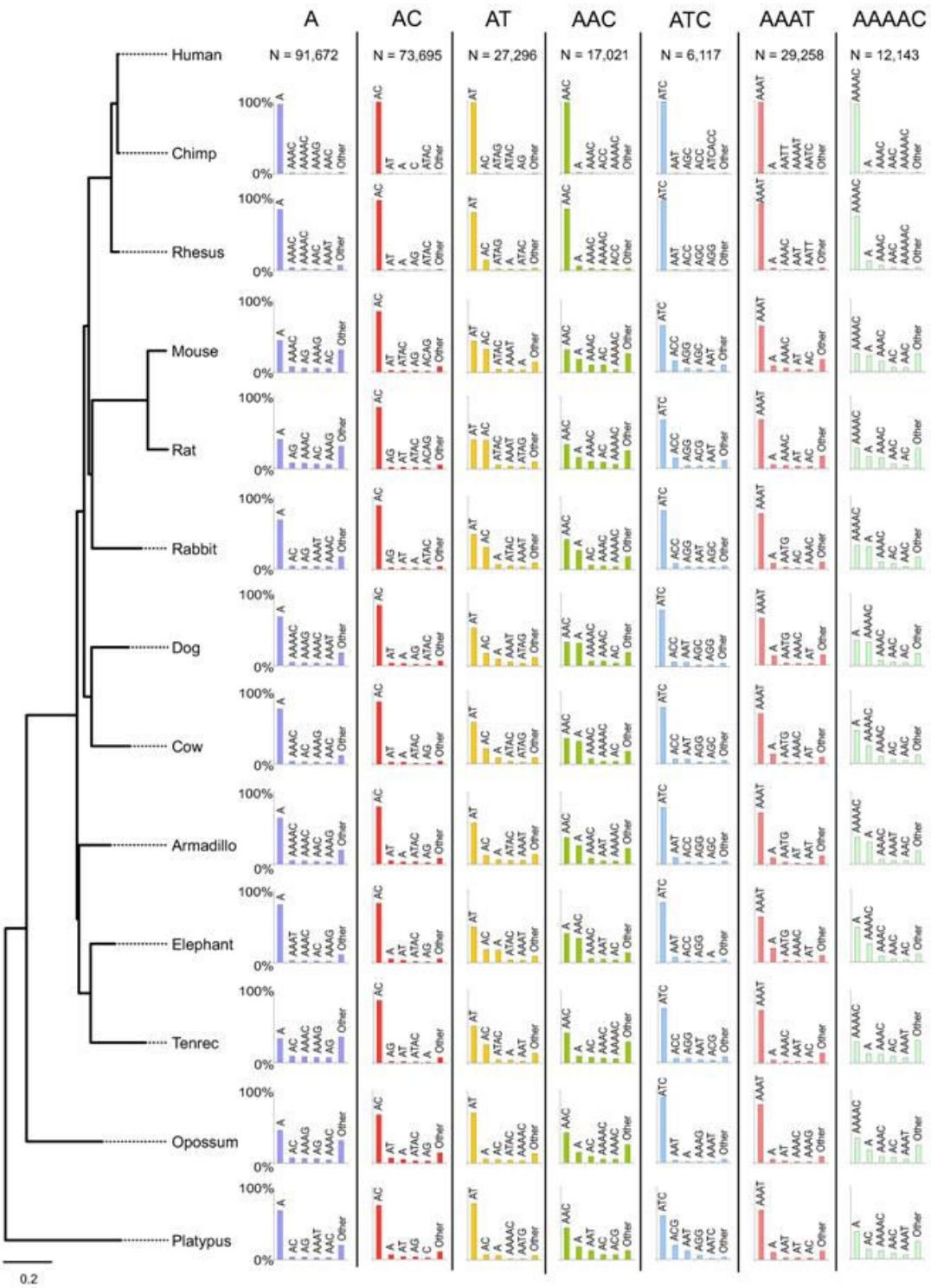


Figure 5.2: Motif replacement of simple human conserved microsatellites in orthologous positions of mammalian genomes. N: number of simple human microsatellites conserved in at least one species for each motif considered. Simple microsatellites orthologous to human loci with motifs are colour-coded as follows: A (blue), AC (red), AT (yellow), AAC (green), ATC (light blue), AAAT (light red) and AAAAC (light green). Bar plots indicate the proportion of orthologous loci consisting of the five most represented motifs and the grouped remaining motifs.

5.4.4 Variation in length

If an interspecies difference between distributions of microsatellite length at equilibrium is detected, distinctive mutational probabilities can be safely assumed to explain these differences. Such variability, probably stemming from variation in the efficiency of the repair machinery (Li 2008), may have consequences on the reliability of cross-species analyses, e.g. comparative studies of co-occurring species at the community level (Whitham et al. 2006). In addition, it may help explain broad patterns of genome evolution, e.g. expansion of the human genome (Kehrer-Sawatzki and Cooper 2007). At the genome scale, human microsatellites have been shown to be on average shorter than mouse (Waterston et al. 2002), rat (Gibbs et al. 2004), opossum and lizard microsatellites, but longer than chicken and platypus microsatellites (Warren et al. 2008). Although informative at the broad scale, these analyses did not compare lengths between orthologous microsatellites, thus incorporate noise in the comparisons. Recent human-chimpanzee comparisons have avoided this bias, but little significant difference could be observed between these two species. Among all classes of microsatellites, only human dinucleotide repeats have been found to be significantly longer than their chimpanzee counterparts (Cooper et al. 1998; Webster et al. 2002; Sainudiin et al. 2004), although Vowles and Amos (2006) claimed, but did not show, that they had found significant differences for all motif size classes. In this analysis, we sought to compare the length of microsatellites conserved in a larger group of mammals to increase the generality of our findings. In addition, we sought to tease out the role of motif type and repeat number on length variation in restricted subsets of species including human, chimpanzee and/or mouse.

First, we examined length variation between orthologous microsatellites ubiquitously conserved in all 13 species examined. Of 63 ubiquitous simple microsatellites with no motif replacement, 38 showed length variability in at least one species. Figure 5.3 depicts the length distribution of these microsatellites in all species. Based on decreasing mean allele length, species can be ranked in the order: rat>platypus>mouse>rhesus>human>opossum>dog>chimpanzee>rabbit>armadillo>cow>tenrec>elephant. However, due to the very low number of loci examined, we could not find statistical support for these differences.

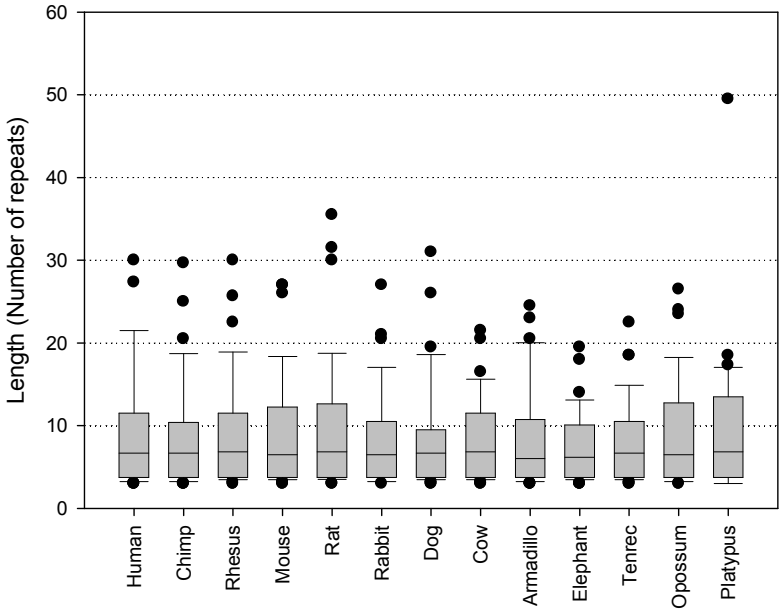


Figure 5.3: Length variation at 22 highly conserved mammalian microsatellites.

To increase the initial number of orthologous microsatellites to compare while maintaining broad coverage of mammalian species, the combination of taxa examined was restricted to the Boreoeutheria (human, chimpanzee, rhesus, mouse, rat, rabbit, dog and cow), i.e. the combination of eight species containing the greatest number of orthologous simple microsatellites of similar motif and variable length (n = 506). Strikingly, mouse and rat orthologues showed a more dispersed distribution of microsatellite length variation (Figure

5.4), and an apparent average expansion of microsatellite sequences, compared to other species. Ranking taxa relative to average microsatellite length was mostly equivalent to the comparison of wide-ranging mammalian microsatellites, giving the order: rat>mouse>dog>human>rhesus>chimpanzee>cow>rabbit. Statistical support for these apparent differences was only found for a fraction of the comparisons (Table 5.5), but still endorses the order: rat=mouse>dog/primates>cow=rabbit (Figure 5.4), with chimpanzee microsatellites significantly shorter than dog and human microsatellites. This classification, though only partially supported, corroborates the few genomewide studies investigating microsatellite length differences between mammalian species: human-chimpanzee comparisons showed that human has on average longer microsatellites than chimpanzee (Sainudiin et al. 2004; Kayser et al. 2006; Vowles and Amos 2006), and mouse microsatellites were reported to be longer than human microsatellites (Waterston et al. 2002) but of similar length to rat microsatellites (Gibbs et al. 2004).

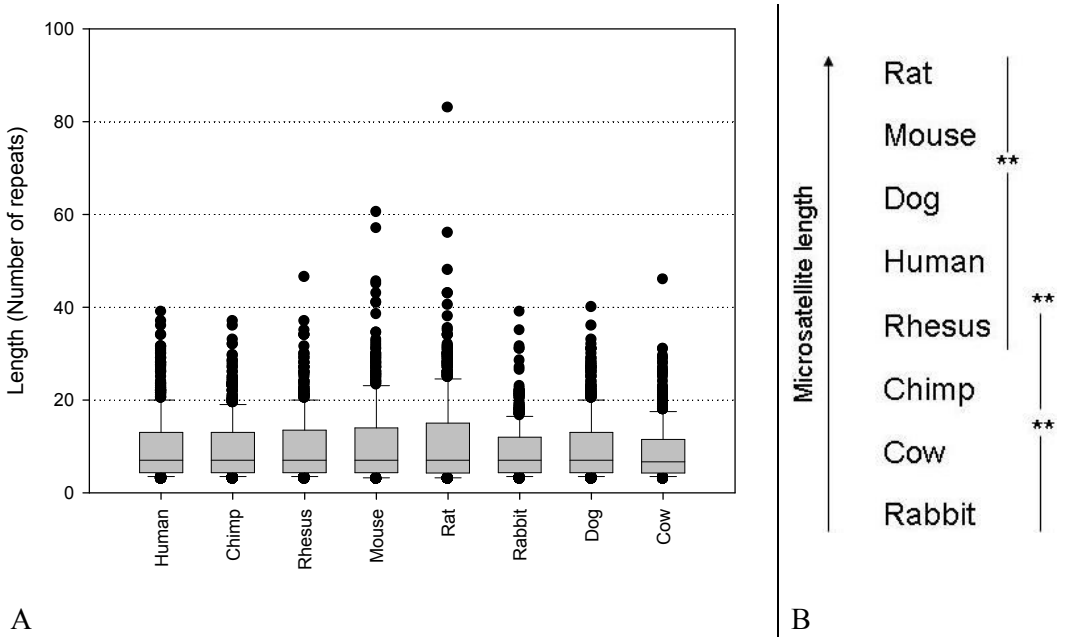


Figure 5.4: Length distribution of 506 orthologous microsatellites in boreoeutherian species. (A) Barplots showing length distribution. (B) Species ranking order relative to average microsatellite length. Lines group species which length distribution could not be differentiated. See Table 5 for statistical analyses. ** indicates P-value, with 0.001<**<0.01.

Table 5.5: Length distribution differences between species pairs. Wilcoxon signed rank test with continuity correction.

Comparison	Alternative Hypothesis	Wilcoxon's V	P-value	Conclusion
Rat-Mouse	Rat>Mouse	37,608	0.2262	n.s.
Mouse-Dog	Mouse>Dog	69,833	0.0082	Mouse>Dog
Dog-Human	Dog>Human	43,215	0.4234	n.s.
Human-Rhesus	Human>Rhesus	25,063.5	0.3057	n.s.
Dog-Rhesus	Dog-Rhesus	43,291	0.1157	n.s.
Rhesus-Chimp	Rhesus>Chimp	26,242	0.1255	n.s.
Dog-Chimp	Dog>Chimp	47,543.5	0.0240	Dog>Chimp
Human-Chimp	Human>Chimp	15,312.5	0.0056	Human>Chimp
Chimp-Cow	Chimp>Cow	53,766	0.0035	Chimp>Cow
Cow-Rabbit	Cow>Rabbit	50,042	0.5996	n.s.

To determine whether these differences subsist in the context of motif type, orthologues found from comparisons of the human, chimpanzee and mouse genomes were binned into groups of identical motif, and empirical cumulative distribution functions (ECDF) were constructed and compared. Figures 5.5 and 5.6 show the ECDFs for major motif classes with confidence intervals. These curves, representing the non-parametric estimates of species-specific stationary distributions, can be used to determine whether the probability to find longer microsatellite is higher in one distribution than in another for ranges of possible lengths. Practically, this difference is illustrated by non-overlapping curves and confidence intervals.

First, considering the total subset of 9,964 orthologous microsatellites, mouse microsatellite length was, on average, greater than that of human microsatellites using non-parametric standard statistics (Wilcoxon signed rank test, $V = 27533553$, $P < 0.0001$), which in turn was longer than that of chimpanzee microsatellites (Wilcoxon signed rank test, $V = 8435249$, $P < 0.0001$), confirming previous findings (Waterston et al. 2002; Sainudiin et al 2004; Kayser et al. 2006; Vowles and Amos 2006). Inspecting the ECDF (Figure 5.5), similar proportions of microsatellites were observed below seven repeats in the three species but, above this value, mouse microsatellites were more likely to be longer

than their primate counterparts. Differences between human and chimpanzee were not as pronounced as mouse-primate comparisons, but the probability of finding longer microsatellites in human than chimpanzee was significant for microsatellites between 14 and 20 repeats. For all three species, few microsatellites expanded over 25-30 repeats, causing confidence intervals to enlarge and overlap between species.

When microsatellites with various motif lengths were considered, all classes of human microsatellites appeared longer than chimpanzee microsatellites and shorter than mouse microsatellites (Figure 5.5). However, differences were only significant (non-overlapping confidence intervals) for dinucleotide microsatellites, which is by far the most common motif found in mammalian genomes. This observation was equivalent to a previous comparison of human-chimpanzee microsatellite length, in which statistical significance was also found for dinucleotide repeats only (Webster et al. 2002).

Among dinucleotide repeats, the majority (86.4%) were composed of $(AC)_n$ microsatellites. Human $(AC)_n$ microsatellites were more likely to be longer than their chimpanzee orthologues in the range 12-22 repeats. In addition, mouse $(AC)_n$ microsatellites were much longer than their primate counterparts; in fact, the probability of finding a longer microsatellite in mouse than in primates was higher for $(AC)_n$ microsatellites between ~7 and 30 repeat units, suggesting that the observed overall pattern of heightened microsatellite length in mouse was greatly influenced by this abundant motif. Interestingly, $(AG)_n$ microsatellites were also likely to be longer than their primate equivalents in the 5-27 repeat length range, despite a fairly restrained sample size (n = 416). In contrast, no difference was detected between human and chimpanzee $(AG)_n$ microsatellites, suggesting that the mutational properties of this type of microsatellites are

quite different in mouse. We did not find any other difference among orthologues of human, chimpanzee and mouse.

To reach optimal statistical power to detect length differences in the context of motif types and repeat number, microsatellites found in human-chimpanzee (n = 176,036) and human-mouse (n = 11,518) comparisons were isolated; we believe these subsets to comprise all pairwise simple microsatellites of identical motifs and variable length.

These enlarged sample sizes enabled us to find significant length differences between human and chimpanzee microsatellites (Figure 5.6): human>chimpanzee: $(AC)_n$ (7-27 repeat range), $(AG)_n$ (12-21 repeat range) and $(AT)_n$ (7-26 repeat range) microsatellites; chimpanzee>human: $(A)_n$ microsatellites (12-21 repeat range). However, we could not differentiate length variation between microsatellites composed of other motifs.

Human-mouse comparisons showed longer mouse microsatellites for motifs A (19-27 repeat range), AC (7-30 repeat range), AG (7-28 repeat range), but not for other motifs. Some primate trinucleotide repeats, e.g. $(CCG)_n$ microsatellites, appeared longer in primates than in mouse in the three-way comparison (data not shown), although the confidence intervals overlapped. This could be of significance as over-expansion of exonic $(GCC)_n$ repeats are responsible for at least six human diseases involving fragile sites (Pearson et al. 2005). However, no difference was observed among human and mouse $(CCG)_n$ orthologous microsatellites, suggesting that either there is no genuine difference or that there might only be a detectable difference in the subset of $(CCG)_n$ microsatellites that are broadly conserved and/or selected upon.

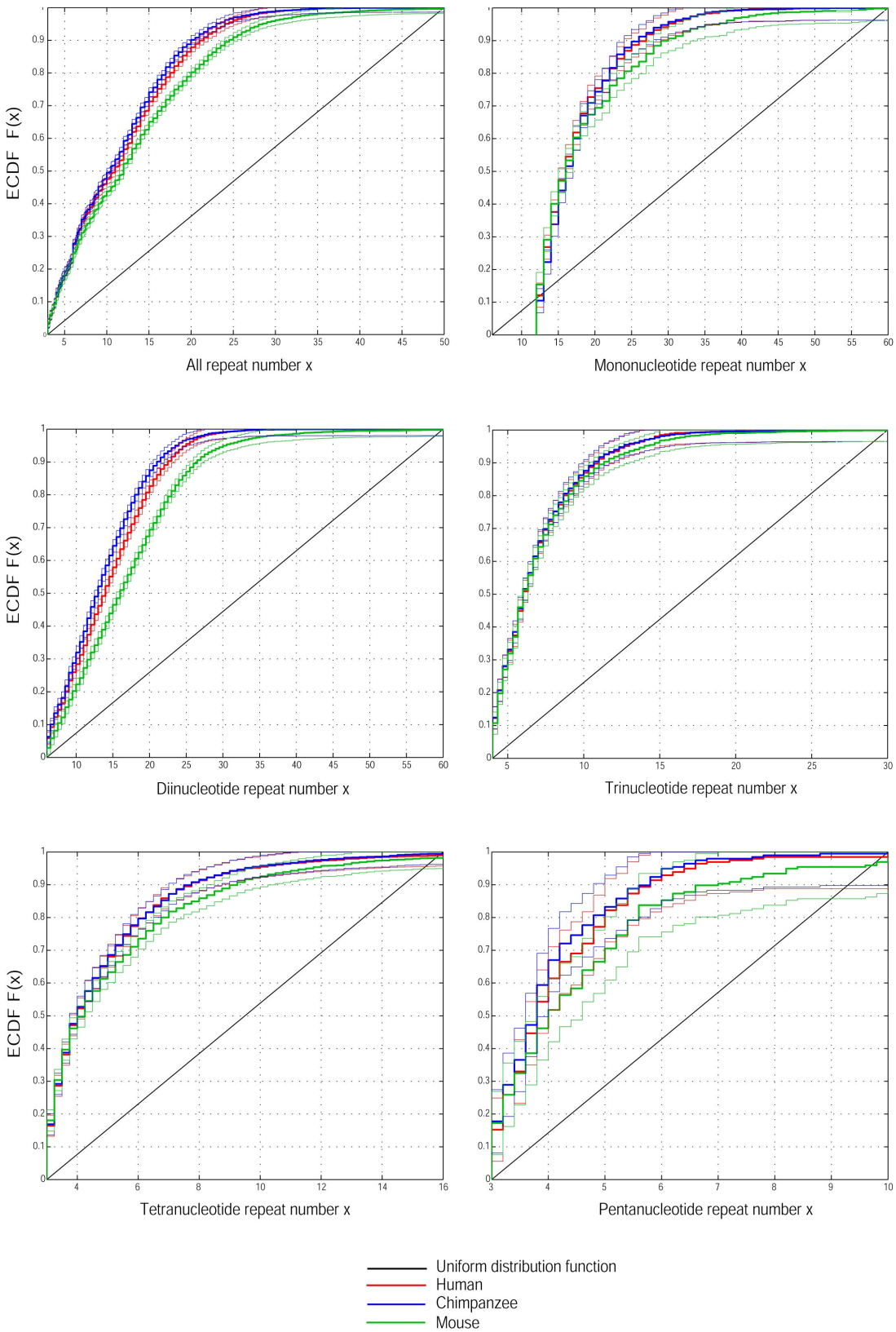


Figure 5.5: Non-parametric estimation of species-specific stationary distribution. Empirical cumulative distribution function (ECDF) of microsatellite lengths (repeat number) in human, chimpanzee and mouse. ECDFs are represented in bold lines, and 95% confidence bands in fine lines.

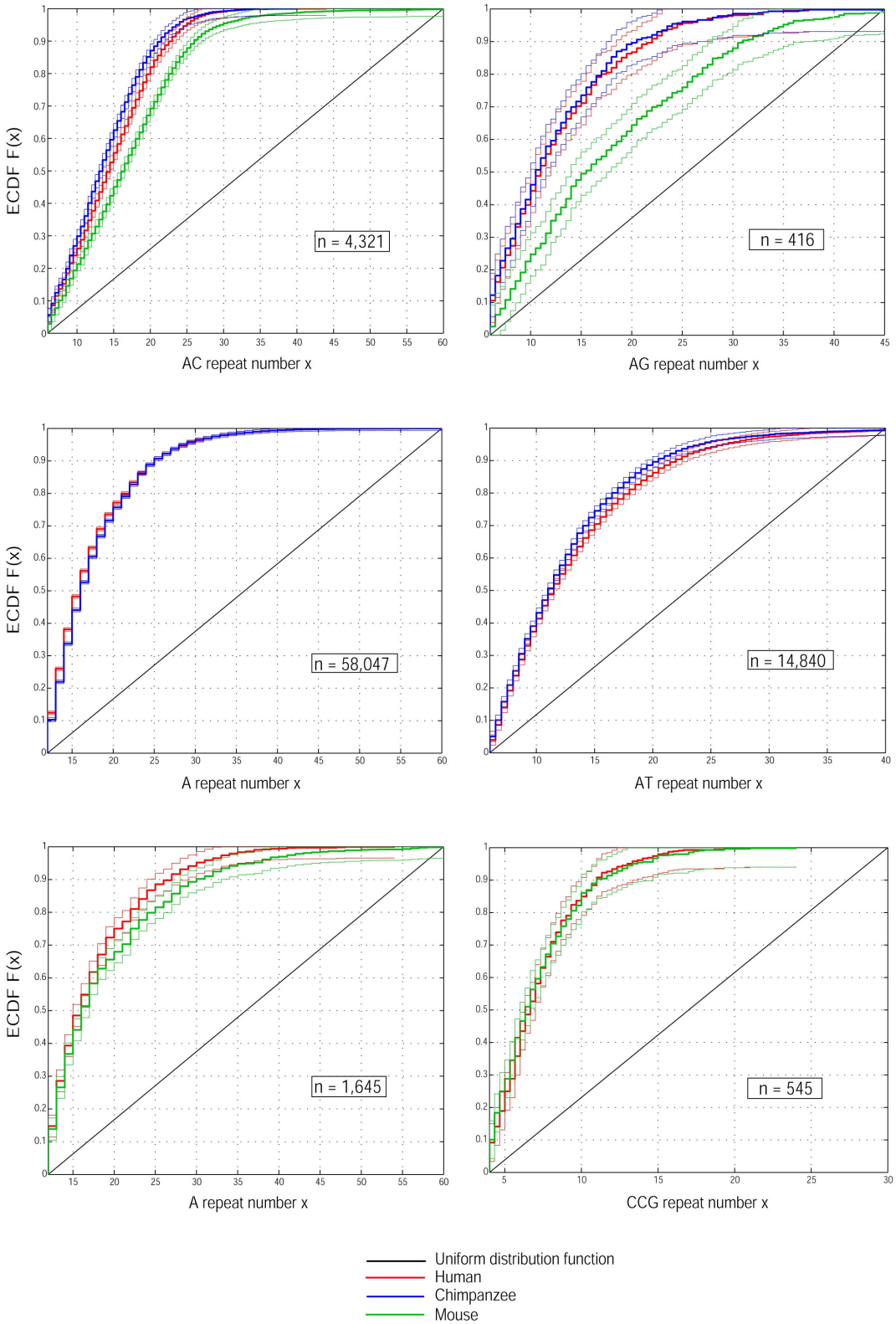


Figure 5.6: Non-parametric estimation of species-specific stationary distribution in the context of motif composition. Empirical cumulative distribution function (ECDF) of microsatellite lengths (repeat number) in human, chimpanzee and/or mouse comparisons and their 95% confidence bands.

5.5 Discussion

In this study, we took advantage of the wealth of data given by the conservation of microsatellites in mammalian genomes to investigate the nature and extent of microsatellite structural changes above the species level. Overall, the pattern was quite striking and often unexpected.

First, we found that there was a high rate of simple (ancestral) → compound (derived) microsatellite changes in each primate lineages, especially in that of human, whereas compound→simple changes were rarer, confirming the view that the human genome is expanding (Kehrer-Sawatzki and Cooper 2007) and that microsatellite births occur more frequently than degeneration and death of the locus. Most compound microsatellites emerged from motifs derived through point mutations involving changes toward C and G, a bias that goes against the documented genomewide CG→AT mutation bias in mammalian genomes (Lipatov et al. 2006).

Second, we reported that not only can new segments derive from a pre-existing simple microsatellite to form a compound microsatellite, they can altogether replace the ancestral segment for the entire length of the array, a drastic change that, unexpectedly, may occur relatively often for certain types of microsatellites. Striking motif-specific features of motif replacement were observed, with some motifs showing exceptional resistance to replacement (e.g. AC, ATC, CCG and AAAT), whereas other motifs could be replaced in a large fraction of orthologous microsatellites (e.g. AT, AAAAC). Comparatively more motif replacement events occurred in lineages with large phylogenetic tree branch lengths, e.g. rodents, implying elevated mutability and/or short generation time and/or low effective population size, and according with previous reports of elevated

change in these divergent genomes (Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005).

Third, we sought to detect differences in microsatellite length distribution between orthologues. Within the Boreoeutheria, rodent microsatellites were found to be on average longer, whereas cow and rabbit, despite the close relationship of the latter to rodents, were shorter. Primate and dog microsatellites showed an intermediary length distribution. ECDFs were constructed as efficient non-parametric estimates of marginal species-specic stationary distribution of repeat lengths from the orthologous set of loci across a given set of species. The associated 95% condence bands for these species-specic stationary distributions rigourously account for the effect of the finite sample size. i.e. the number of orthologous loci for the class of microsatellites under consideration. ECDFs have the additional advantage of revealing differences within length ranges. Apparent variations in average microsatellite lengths in the context of motif class and composition were observed, although the asymmetry was often statistically insignificant, except for the more numerous mononucleotide and dinucleotide repeats, mostly confirming the order: mouse>human>chimpanzee.

Altogether, these results greatly enhance our knowledge of the evolutionary dynamics of microsatellites in mammalian genomes, and in particular emphasize the interaction of species-specific and motif-specific influences. If species-specific factors (e.g. efficiency of mismatch repair) affected microsatellite mutability on top of motif-specific effects (e.g. motif length and nucleotide composition), we would expect to find consistent differences between species, irrespective of motif type. In contrast, if no pattern common to all species appeared, species-specific factors might be considered insignificant compared to motif type influence. Our results demonstrate that both influences can determine the fate of

microsatellites above the species level, at different levels for different motifs and in different species, confirming that microsatellite evolution is a highly dynamic and heterogeneous process (Ellegren 2004; Buschiazzo and Gemmell 2006).

Comparing length distribution in several mammalian genomes are useful to reveal such differences, but this approach lacks some power because the marginal species-specific distribution lacks the information in the joint distribution from the orthologous data. The stationary distribution inferred from full genome data (Calabrese and Durrett 2003; Sainudiin et al. 2004) also lacks power in comparison to using all the information in the joint distribution at orthologous loci. The marginal distribution from our orthologous set of loci has the initial ascertainment bias of starting the search from humans that becomes pronounced with evolutionary distance, in terms of limiting the sample sizes. However, our non-parametric confidence bands account for the sample size effect. On average, length variation at orthologous mammalian microsatellites demonstrated that significant differences exist between species, confirming previous results in mammals (Cooper et al. 1998; Crawford et al. 1998; Waterston et al. 2002; Webster et al. 2002; Vowles and Amos 2006; Laidlaw et al. 2007), invertebrates (Ross et al. 2003) and plants (Azaiez et al. 2006).

Lastly yet importantly, our investigation allowed the identification of the most appropriate types of microsatellite markers for cross-species transfer in mammalian species. Ideally, such markers would be present in large numbers to facilitate the isolation of a subset of likely polymorphic microsatellite sequences, i.e. long and pure repeat tracts (Pardi et al. 2005). In addition, there should not be more than expected structural change (simple→compound changes, motif replacement) to prevent using markers exhibiting distinct mutational dynamics between species. Interestingly, $(AC)_n$ microsatellites, which are already widely used as genetic markers, meet these requirements. In addition, the less

utilized $(ATC)_n$ and $(AAAT)_n$ microsatellites stood out as other potential markers and their development for cross-species applications should therefore be promoted.

5.6 Acknowledgments

R. Sainudiin inspired the use of ECDFs to detect length differences between species and provided both MATLAB scripts and advice to improve the analysis.

Chapter 6

6 General Summary and Conclusion

6.1 Background

Microsatellite DNA, a type of variable simple sequence repeat, represents ~3% of the mammalian genome (Warren et al. 2008). Despite uncertainties regarding its functionality and mutational dynamics (Ellegren 2004; Buschiazzo and Gemmell 2006), various fields of research take advantage of the fraction of genomic microsatellites that exhibits extreme variability (Pardi et al. 2005), e.g. to discriminate individuals within a population (Blouin 2003; Butler 2005), populations within a species (Sunnucks 2000) and, to a certain extent, among species (Mikul et al. 2007). Substantial efforts have been made to improve our knowledge of how microsatellites evolve (Vargas-Jentzsch et al 2008), both in the short term (through, e.g., pedigree analyses and sperm typing) and in the long term (using, e.g., phylogenetic inference and genome comparisons). This would enable the development of more realistic models of evolution (Ellegren 2004; Buschiazzo and Gemmell 2006). However, short-term approaches are limited in their evolutionary scope and long-term analyses are often too anecdotal to obtain a comprehensive picture of the evolution of microsatellites above the species level. Genome comparisons have recently bridged the gap, but still lacked polymorphism data, a prerequisite to inspect ongoing mutational processes (but see Brandström and Ellegren 2008). In addition, genome-scale comparisons have been, in mammals, restricted to human-chimpanzee analyses (Vowles and Amos 2006; Kelkar et al. 2008), species that diverged only ~6-7 Myr ago (Steiper and Young 2006). An approach that encompasses a long evolutionary period and accounts for intraspecific variability is therefore needed, yet critically relies on the identification of microsatellites conserved above the species level.

6.2 Overview

Revelations that some microsatellites may be conserved for 100+ Myr (FitzSimmons et al. 1995; Rico et al. 1996; Moore et al. 1998) challenged the general assumption that microsatellite sequences are too labile to be retained in genomes over a large evolutionary scale, and suggest that many more remain to be discovered. In this thesis, I sought to identify human microsatellites conserved in 17 vertebrate genomes, including those of a marsupial and a monotreme, using approaches in comparative genomics and the wealth of data created by recent sequencing projects. Drawing on this comprehensive dataset, I have been able to determine whether microsatellite conservation was mostly driven by neutral forces (Chapter 2), and further investigated how conserved microsatellites may be applied for phylogenetic reconstruction (Chapter 3), to develop PCR primers transferable across the Mammalia (Chapter 4) and to study the nature and amount of microsatellite structural change above the species level (Chapter 5). I anticipate my results to open new windows on how the presence of microsatellites in genomes is perceived and how we may take advantage of microsatellite retention to study their evolution and to transfer microsatellite markers across species for various applications, e.g. in comparative mapping and population genetics.

6.3 Is the retention of microsatellites in vertebrate genomes driven by neutral forces only?

Methods of comparative genomics have proven efficient to find widespread active conservation in mammalian non-coding sequences (Waterston et al. 2002; Smith et al. 2004; Lindblad-Toh et al. 2005; Siepel et al. 2005; Venkatesh et al. 2006; Mikkelsen et al.

2007), a feature that was unexpected before the rise of sequencing projects (Dermitzakis et al. 2005). As for microsatellite sequences, whose function in the genome is unclear and mutation rates are several orders of magnitude higher than the genomic average (Buschiazzo and Gemmell 2006), they are expected to be cleared from genomes rapidly, or else, maintained only by chance. Growing lines of evidence indicate that microsatellites appear to evolve following a life cycle pattern, from birth, through expansion, contraction and finally death (Figure 1.2), yet their life expectancy above the species level is quantitatively unknown (Stephan and Kim 1998).

To elucidate the extent of conservation of microsatellites in vertebrate genomes, I used whole-genome alignments and identified 594,340 human microsatellites conserved in at least one of 11 mammalian and five non-mammalian vertebrate genomes. Conservation of microsatellites appeared to decrease exponentially with increasing evolutionary time (Figure 2.3), an indication that the decay process of most vertebrate microsatellites is probably a consequence of genetic drift. Nevertheless, a significant fraction of human microsatellites was found in large subsets of mammalian species, and even through to amphibian, avian and fish species. This finding suggested either that at least some loci have been actively selected for because they serve an advantageous biological function, that they have been maintained passively in regions of the genome under strong selection, or that they lie in regions that stochastically experienced fewer substitutions than others, even in the absence of selective effects.

A multiple whole-genome alignment assigns homology between nucleotides, but it does not identify genomic positions that are under selection or evolving neutrally, unless a neutral model of evolution can be applied to differentiate among these. Unfortunately, there are currently unresolved theoretical issues regarding the development of a null expectation applied to microsatellite sequences to distinguish between mere retention and

active conservation. In view of growing evidence of microsatellite functionality (Li et al. 2004; Kashi and King 2006), and because conservation is only one of the attributes of cis-regulatory elements and is neither necessary nor sufficient (Vardhanabhuti et al. 2007), more work is therefore needed to examine the possible role of active selection into shaping the intricate patterns of microsatellite conservation in vertebrate genomes. Such advances may allow biologists to identify and validate experimentally the subset of functionally conserved microsatellites, if any.

6.4 Are microsatellite presence/absence data suitable for phylogenetic reconstruction?

Microsatellite variability has been used in conjunction with genetic distance methods (Goldstein et al. 1995; Shriver et al. 1995; Slatkin 1995) to reconstruct shallow phylogenies (Bowcock et al. 1994; Meyer et al. 1995; Paszek et al. 1998; Richard and Thorpe 2001; Ayub et al. 2003; Mikul et al. 2007; Rout et al. 2008). However, issues stemming from the heterogeneous mutational dynamics of microsatellites above the species level, e.g. upper allele constraints and size homoplasy (Noor et al. 2001; Estoup et al. 2002), have hampered their use to infer deeper relationships. Investigators could only rely on information found in the flanking sequences to produce reliable species trees (Makova et al. 2000; Martin et al. 2002; Domingo-Roura et al. 2005; Shepherd and Lambert 2005). Imperfections in the repeat array also appeared to be phylogenetically informative (Zhu et al. 2000).

Drawing on approaches recently used to construct phylogenies from gene content (Huson and Steel 2004), I sought to take advantage of the presence/absence data of microsatellite

sequences in vertebrate genomes to reconstruct and assess a vertebrate phylogeny, including monotremes, using models of evolution optimized for alternative data (i.e. restriction sites and morphological characters). To this end, the information regarding microsatellites conserved in platypus sequences of a 6-way alignment (Chapter 3) was integrated into the framework created by the human 17-way alignment analysis (Chapter 2). Overall, maximum parsimony (MP) analyses using the tree-bisection-reconnection (TBR) branch-swapping algorithm (Swofford 2002) resulted in a tree more similar to the current consensus tree (Bininda-Emonds et al. 2007; Miller et al. 2007) than Bayesian analyses assuming either a morphological model with variable rates of gain and loss (Pagel and Meade 2004) or a restriction site model with equal rates of change (Huelsenbeck and Ronquist 2001). In fact, the MP tree topology was almost identical to the authoritative tree, including the position of platypus at the base of the Mammalia, but with rodents misplaced at the base of the Euarchontoglires (Chapter 3, Figure 7).

Phylogenetic reconstruction based on microsatellite presence/absence transitions is therefore potentially adequate, at least in vertebrate species, but further theoretical developments are needed to account for the specific evolution and persistence of microsatellite sequences. Such developments may provide opportunities to improve phylogenetic inference from the current dataset of conserved microsatellites and to reconstruct alternative phylogenies once additional genomes are incorporated into the existing framework. In addition, neutral expectations may be tested using a sliding window approach to sequentially reconstruct phylogenetic trees from microsatellite conservation/absence and find regions that deviate from the genomewide, presumably neutral retention pattern, and result in divergent tree shapes.

6.5 Can conserved microsatellite primers be transferred across the Mammalia?

Microsatellites are currently one of the most popular types of genetic markers for molecular ecology (Sunnucks 2000; Rossiter et al. 2007) and genome mapping studies (Weissenbach et al. 1992; Luo et al. 2007). However, their applications could be facilitated, or even extended, if PCR primers could be readily transferred between species. Although numerous cases of cross-species transfers have been reported (e.g. Gemmell et al. 1997; Guillemaud et al. 2000; Kim et al. 2004; MacDonald et al. 2006; Schlötterer et al. 1991), the transferability of microsatellite markers decreases critically with increasing time of divergence between species, and only in rare occasions has it been demonstrated in highly distant species, i.e. species that diverged 100+ Myr ago (FitzSimmons et al. 1995; Rico et al. 1996; Moore et al. 1998).

The large extent of conservation that I found among mammalian genomes, including a marsupial and a monotreme species (Chapter 2 and 3), and the likelihood that a fraction of these may lie in regions of low mutability, suggested that at least some microsatellites may be retained in mutation-purified sequences prone for the design of cross-species primers. Out of a random set of ~1000 broadly conserved dinucleotide repeats, 19 loci allowed the design of degenerate comparative primers, nine of which were suitable for cross-species amplification and M13-genotyping (Schuelke 2000) in most of the 18 species included in the study. In addition, the five most successfully genotyped loci were directly sequenced in homozygote individuals to evaluate the sequence structure and the nature of the mutations leading to any allele length polymorphism observed through genotyping. Drawing on this limited sequencing effort, I estimated that 75% of the length variability detected was consistent with the occurrence of stepwise forward/backward mutations of the longest pure

repeat tract forming the microsatellite, a parameter that can be easily modeled under currently implemented models of evolution.

Overall, the evidence suggests that, out of an initial random set of ~1000 conserved dinucleotide repeats, three loci are polymorphic in most if not all mammalian species, and are highly suitable as genetic markers for comparative analyses. By extrapolation, I estimated that at least nine other dinucleotide repeats could be identified and applied for cross-species applications across the Mammalia, but this number could increase significantly using less stringent selection criteria or if other types of microsatellites than dinucleotide repeats are considered.

6.6 What are the nature, extent and consequences of structural change in microsatellite DNA above the species level?

Ideally, cross-species microsatellite markers should exhibit similar mutational dynamics between species to fully implement the comparative approach. Microsatellite mutability has been shown to depend mostly on the internal structure of the repeat array (e.g. array length and purity, and motif length and composition), but selective influences may also add to the heterogeneity of microsatellite evolution, e.g. interspecies variability in the efficiency of the repair machinery (reviewed in Buschiazzo and Gemmell 2006). With only few reports investigating microsatellite structure evolution in closely related species at a few loci only (e.g. Zhu et al. 2000), there was a need for a genome-scale analysis of structural change at orthologous loci.

First, I described in Chapter 1 how A+T-rich microsatellites were the most abundant class of microsatellites in human sequences, but also disappeared more rapidly than other

microsatellites, suggesting that the turnover of A+T-rich microsatellites is much higher than that of G+C-microsatellites in mammalian genomes. This first indication that microsatellites consisting of different motifs evolve differentially was consistent with a heightened mutability of A+T-rich microsatellites in primate genomes (Kelkar et al. 2008). In addition, I detailed in Chapter 5 the unexpected extent of structural change that can be observed among orthologous microsatellites, including: (i) in primates, a higher rate of change from a simple to a compound structure than the reverse change, with most derived motifs originating from base substitutions towards C or G, (ii) motif replacements occuring in large proportions of orthologous microsatellites, with certain motifs and species more prone than others to show elevated rates of motif replacement, (ii) significant differences in the length distribution of large sets of orthologous microsatellites can be detected for some motif types. Overall, these results suggested that there may be an intricate interplay of motif-specific and species-specific factors influencing the mutability of microsatellites.

The exceptional stability and high abundance of $(AC)_n$ microsatellites in mammalian genomes encourages their promotion for comparative marker development; indeed, transferable microsatellites with similar mutational dynamics should be favored against microsatellites showing heterogeneous properties between species. Alternatively, less frequently employed $(ATC)_n$ and $(AAAT)_n$ microsatellites showed stable features similar to $(AC)_n$ microsatellites and may also be adequate for marker development.

6.7 Final comments

In this thesis, I presented the first comprehensive identification of human microsatellites conserved in vertebrate genomes and reported a number of insights into the implications and applications of such deep microsatellite conservation, including that (i) the persistence

of microsatellites in genomes is predominantly a random process, but at least some widely conserved microsatellites may be actively selected for (ii) the presence/absence state of microsatellites can be used to reconstruct deep phylogenies, (iii) cross-species primers can be developed and transferred for comparative analyses across the Mammalia and (iv) that the amount and nature of structural change is a function of motif-specific and species-specific factors.

These are only a few of the many developments that can be drawn from such a rich dataset, and I anticipate that, if made publicly available, this dataset would prove useful in different fields of research, e.g. comparative mapping, molecular ecology, and models of microsatellite and genome evolution. In particular, non-parameteric Markov models of microsatellite evolution over state spaces that span the pure, interrupted, compound repeats and dead repeats could be built to avoid ascertainment bias. In addition, I suspect that the availability of robust comparative primers, allied with controlled sampling, analysis of length and sequence variability, and the future development of novel statistical approaches to reconstruct ancestor allele states at microsatellite loci in highly divergent species with realistic models of microsatellite mutation over appropriate ancestral histories would promote the use of comparative methods such as independent contrasts to study the evolution and history of microsatellites above the species level (Zhu et al. 2000).

7 Appendix

7.1 Table appendix

Table 1: Microsatellites in the 17-WA. Perf: perfect microsatellites; imperf: imperfect microsatellites; 1x: mono-, 2x: di-, 3x: tri-, 4x: tetra-, 5x: penta-, 6x: hexanucleotide simple repeats; comp: compound; conserved: number of microsatellites conserved in at least one species (for human), or in human (for non-human species).

	Simple														
Species	Simple perf	Simple imperf	1x	2x	3x	4x	5x	6x	Total simple	Total Comp	Total linked	Total Mixed	Total number	% Human	Conserved	% Human
Human	429,184	191,300	119,640	175,976	96,975	174,720	43,083	10,090	620,484	44,000	24,583	10,403	696,016	100%	594,340	100%
Chimp	387,081	166,048	115,611	154,997	83,400	153,984	36,329	8,808	553,129	32,511	19,447	6,579	611,666	87.88%	521,476	87.74%
Rhesus	365,353	146,809	134,897	131,501	71,588	134,817	30,880	8,479	512,162	29,202	15,948	4,889	562,201	80.77%	332,874	56.01%
Mouse	234,973	88,508	81,118	82,173	40,913	86,945	24,761	7,571	323,481	22,224	8,705	1,968	356,378	51.20%	42,633	7.17%
Rat	175,839	96,538	47,687	71,661	31,683	70,983	17,444	5,512	244,970	19,014	6,784	1,609	272,377	39.13%	37,018	6.23%
Rabbit	176,235	65,941	69,132	57,875	29,555	55,937	12,364	4,726	229,589	7,990	4,100	497	242,176	34.79%	40,452	6.81%
Dog	330,279	119,867	164,481	85,371	55,015	103,504	30,783	10,992	450,146	18,607	9,626	1,335	479,714	68.92%	74,042	12.46%
Cow	264,801	73,514	118,981	73,368	43,228	81,540	17,735	3,463	338,315	8,803	5,715	575	353,408	50.78%	57,718	9.71%
Armadillo	150,768	46,982	52,538	38,527	29,392	58,918	14,632	3,743	197,750	4,789	3,125	292	205,956	29.59%	32,973	5.55%
Elephant	261,604	61,490	155,358	52,548	36,036	63,690	13,001	2,461	323,094	8,351	5,446	489	337,380	48.47%	45,638	7.68%
Tenrec	131,730	35,449	26,530	44,085	29,313	51,859	12,468	3,624	167,179	5,761	3,425	337	176,702	25.39%	22,203	3.74%
Opossum	73,182	27,816	20,582	23,614	19,129	37,068	8,003	2,592	100,998	4,143	2,394	517	108,052	15.52%	10,140	1.71%
Chicken	18,920	4,944	10,546	1,811	4,404	5,697	1,177	209	23,864	394	347	25	24,610	3.54%	1,965	0.33%
Frog	10,563	3,370	6,554	2,253	2,446	2,269	339	72	13,933	361	278	49	14,621	2.10%	1,328	0.22%
Zebrafish	11,065	4,021	2,952	4,207	3,554	3,447	705	134	15,086	993	643	245	16,880	2.43%	1,824	0.31%
Fugu	5,809	2,387	1,526	2,033	2,929	1,260	256	192	8196	349	217	46	8,808	1.27%	961	0.16%
Tetraodon	9,294	4,009	2,490	3,605	4,586	1,683	368	457	13,303	655	373	57	14,274	2.05%	1,322	0.22%

Table 2: Single-copy conserved microsatellites in the human genome. MSATs: all human microsatellites in the alignments; HCM: human conserved microsatellites; PSMs: primate-specific microsatellites; NPM: microsatellites conserved in non-primate species; NP3Ms: microsatellites conserved in at least three non-primate species.

Chr	Unique aligned	MSAT	Density All	Density Unique	HCM	%	NPM	%	NPM_3	%
1	37.27%	54954	244.24	655.37	47157	85.81%	16445	29.93%	3816	8.09%
2	40.43%	59394	249.86	617.92	51353	86.46%	17688	29.78%	3739	7.28%
3	39.20%	48598	249.60	636.77	42070	86.57%	14742	30.33%	2967	5.98%
4	39.03%	46254	246.96	632.81	39863	86.18%	13014	28.14%	2368	5.12%
5	38.96%	43558	245.12	629.23	37711	86.58%	12893	29.60%	2553	5.86%
6	40.85%	41666	249.09	609.74	35918	86.20%	11621	27.89%	2160	5.18%
7	37.65%	36928	238.32	632.99	31380	84.98%	9879	26.75%	1825	4.94%
8	38.59%	36236	254.09	658.35	31074	85.75%	9997	27.59%	1957	5.40%
9	35.09%	27452	228.49	651.08	23606	85.99%	8423	30.68%	1846	6.72%
10	38.71%	32335	245.66	634.65	27843	86.11%	9303	28.77%	1934	5.98%
11	36.90%	31634	241.24	653.73	27183	85.93%	9429	29.81%	2123	6.71%
12	37.75%	33543	257.42	681.83	28763	85.75%	9581	28.56%	2062	6.15%
13	41.72%	24276	254.04	608.98	20929	86.21%	6520	26.86%	1093	4.50%
14	38.99%	21974	248.88	638.40	18938	86.18%	6490	29.53%	1409	6.41%
15	36.67%	18474	227.12	619.37	15853	85.81%	5528	29.92%	1308	7.08%
16	34.17%	20712	262.56	768.38	17534	84.66%	6308	30.46%	1482	7.16%
17	36.98%	19281	247.83	670.21	16268	84.37%	5657	29.34%	1496	7.76%
18	42.26%	19538	261.71	619.20	16922	86.61%	5540	28.36%	1150	5.89%
19	30.14%	14926	267.56	887.72	11508	77.10%	3294	22.07%	853	5.71%
20	36.13%	15693	263.72	729.86	13392	85.34%	4604	29.34%	989	6.30%
21	40.51%	8921	261.06	644.37	7519	84.28%	2282	25.58%	368	4.13%
22	33.43%	8378	240.39	719.08	6881	82.13%	2205	26.32%	458	5.47%
X	28.42%	31291	207.14	728.90	24675	78.86%	7960	25.44%	1652	5.28%
Total	37.41%	696016	243.53	651.00	594340	85.39%	199403	28.65%	41608	5.98%
Y	13.51%	3454	134.64	996.41	990	28.66%	43	1.24%	0	0.00%

Table 3: Statistical correlations between various genomic features in 1 Mb-windows. Left to right: G+C content, gene density, LINE and LTR coverage, indel-puried sequence coverage, average recombina5tion rate, SINE coverage, SNP density and density of conserved transcription factor binding sites. Source: UCSC Genome Browser.

	GC	gene	LINE	LTR	cIND	R_{recomb}	SINE	SNP	tfbs
GC	-	0.69	-0.70	-0.45	0.05	0.38	0.78	0.16	0.28
Gene	0.69	-	-0.48	-0.44	0.03	0.11	0.71	0.02	0.21
LINE	-0.70	-0.48	-	0.49	-0.22	-0.40	-0.68	-0.08	-0.41
LTR	-0.45	-0.44	0.49	-	-0.39	-0.16	-0.52	0.14	-0.53
CIND	0.38	0.03	-0.45	-0.54	-	0.10	0.42	-0.18	0.87
R_{recomb}	0.38	0.11	-0.40	-0.16	0.09	-	0.24	0.35	0.14
SINE	0.78	0.71	-0.68	-0.52	0.10	0.24	-	0.00	0.30
SNP	0.16	0.02	-0.08	0.14	-0.23	0.35	0.00	-	-0.24
Tfbs	0.28	0.21	-0.41	-0.53	0.89	0.14	0.30	-0.24	-

Table 4: IUPAC base nomenclature

IUPAC code					
Symbol	Description	Bases represented			
A	Adenosine	A			
C	Cytidine		C		
G	Guanine			G	
T	Thymidine				T
U	Uridine				U
W	Weak	A			T
S	Strong		C	G	
M	aMino	A	C		
K	Keto			G	T
R	Purine	A		G	
Y	pYrimidine		C		T
B	not A		C	G	T
D	not C	A		G	T
H	not G	A	C		T
V	not T	A	C	G	
N	aNy base	A	C	G	T

7.2 Figure appendix

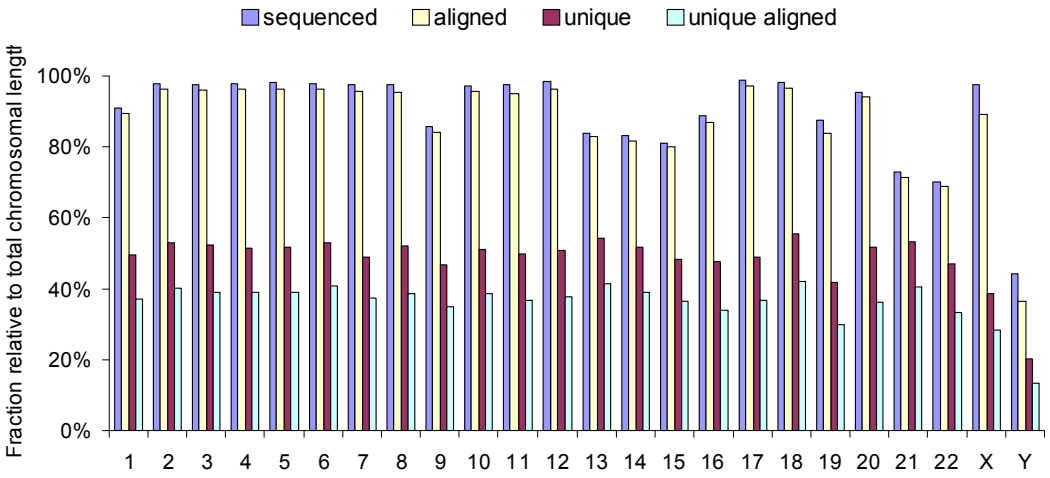


Figure 1: Human chromosomes. Bar plots show the fraction (%) of the estimated size of each chromosome that is sequenced (blue), aligned (yellow), duplication- and repeat-free, i.e. 'unique' (red), and unique aligned (light blue). Source: UCSC Genome Browser.

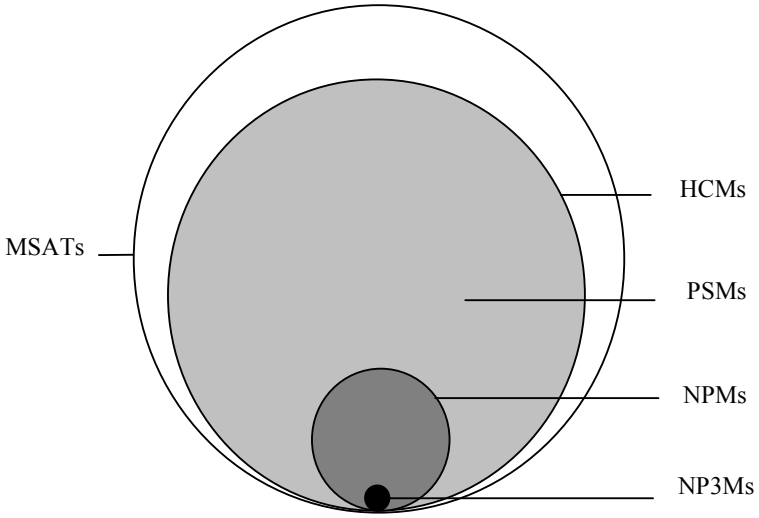


Figure 7: Schematic representation of conserved microsatellite datasets used for statistical analyses. MSATs: all human microsatellites in the 17-WA; HCM: human conserved microsatellites; PSMs: primate-specific microsatellites; NPM: microsatellites conserved in non-primate species; NP3Ms: microsatellites conserved in at least three non-primate species.

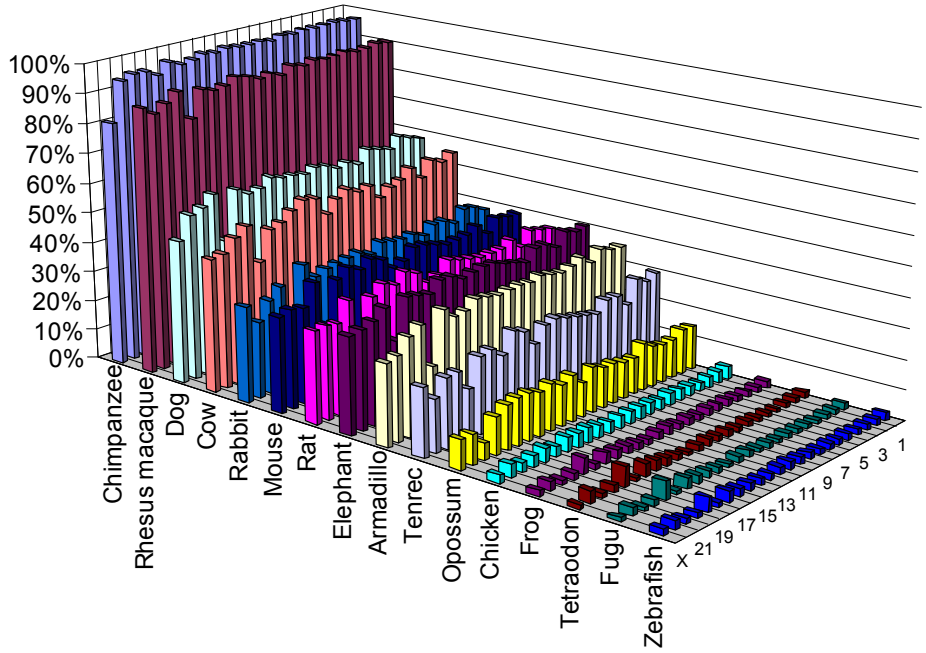


Figure 3: Alignability of human chromosomes with genomes included in the 17-WA (proportion of ungapped length aligned).

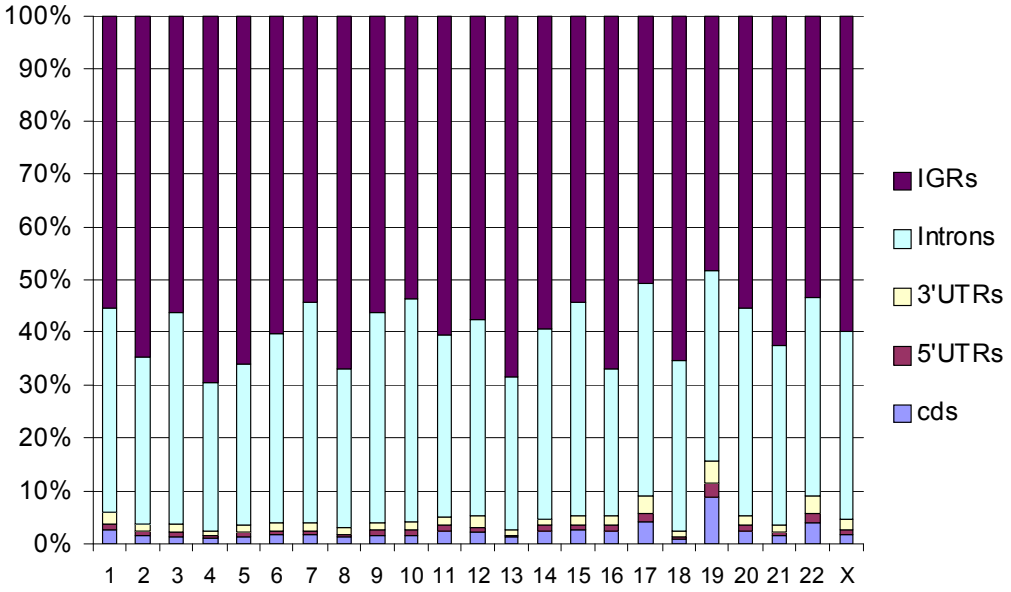


Figure 4: Genomic location of conserved microsatellites in human chromosomes. Primates excluded.

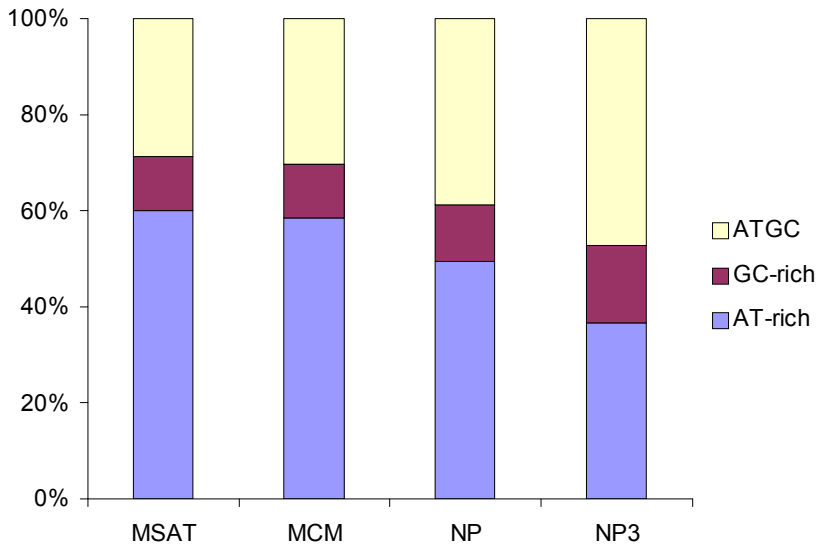


Figure 5: G+C enrichment of conserved microsatellites.

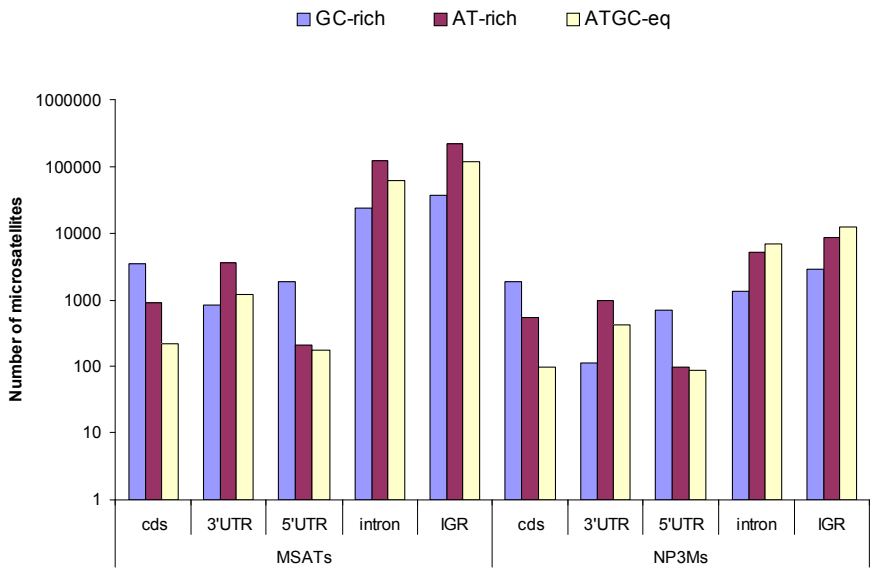


Figure 6: Location of microsatellites relative to their G+C enrichment.

8 References

Ackermann M and Chao L (2006) DNA sequences shaped by selection for stability. PLoS Genetics 2 (2): e22

Adams RI et al. (2004) The impact of microsatellite electromorph size homoplasy on multilocus population structure estimates in a tropical tree (*Corythophora alta*) and an anadromous fish (*Morone saxatilis*). Molecular Ecology 13 (9): 2579-88

Ahituv N et al. (2007) Deletion of ultraconserved elements yields viable mice. PLoS Biology 5 (9): e234

Almeida P and Penha-Goncalves C (2004) Long perfect dinucleotide repeats are typical of vertebrates, show motif preferences and size convergence. Molecular Biology and Evolution 21 (7): 1226-1233

Altekar G et al. (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20 (3): 407-15

Amos B et al. (1993) Social structure of pilot whales revealed by analytical DNA profiling. Science 260 (5108): 670-2

Amos W (1999) A comparative approach to study the evolution of microsatellites. *In* D. Goldstein and C. Schlötterer, Oxford University Press, New York

Amos W and Rubinstzein DC (1996) Microsatellites are subject to directional evolution. Nature Genetics 12 (1): 13-4

Amos W et al. (1996) Microsatellites show mutational bias and heterozygote instability. Nature Genetics 13 (4): 390-1

Anderson TJ et al. (2000) Complex mutations in a high proportion of microsatellite loci from the protozoan parasite *Plasmodium falciparum*. Molecular Ecology 9 (10): 1599-608

Angers B and Bernatchez L (1997) Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Molecular Biology and Evolution 14 (3): 230-8

Angers B et al. (2000) Microsatellite size homoplasy, SSCP, and population structure: a case study in the freshwater snail *Bulinus truncatus*. Molecular Biology and Evolution 17 (12): 1926-32

Arcot SS et al. (1995) *Alu* repeats: a source for the genesis of primate microsatellites. Genomics 29 (1): 136-44

Ayub Q et al. (2003) Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. American Journal of Physical Anthropology 122 (3): 259-268

Azaiez A et al. (2006) Length, orientation, and plant host influence the mutation frequency in microsatellites. Genome 49 (11): 1366-1373

Bachmann L et al. (2004) Allelic variation, fragment length analyses and population genetic models: a case study on *Drosophila* microsatellites. Journal of Zoological Systematics 42 (3): 215-223

Bachtrog D et al. (2000) Microsatellite variability differs between dinucleotide repeat motifs - Evidence from *Drosophila melanogaster*. Molecular Biology and Evolution 17 (9): 1277-1285

Bacon AL et al. (2000) Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. Human Molecular Genetics 9 (18): 2707-13

Baer CF et al. (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. Nature Reviews Genetics 8 (8): 619-631

Bagshaw AT et al. (2008) High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. BMC Genomics 9 (1): 49

Bailey JA et al. (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Research 11 (6): 1005-1017

Barbara T et al. (2007) Cross-species transfer of nuclear microsatellite markers: potential and limitations. Molecular Ecology 16 (18): 3759-3767

Bashir A et al. (2005) Orthologous repeats and mammalian phylogenetic inference. Genome Research 15 (7): 998-1006

Batzer MA and Deininger PL (2002) Alu repeats and human genomic diversity. Nature Reviews Genetics 3 (5): 370-9

Bernardi G (2007) The neoselectionist theory of genome evolution. Proceedings of the National Academy of Sciences of the United States of America 104 (20): 8385-90

Bininda-Emonds ORP et al. (2007) The delayed rise of present-day mammals. Nature 446 (7135): 507-512

Blair JE and Hedges SB (2005) Molecular phylogeny and divergence times of deuterostome animals. Molecular Biology and Evolution 22 (11): 2275-2284

Blanquer-Maumont A and Crouauroy B (1995) Polymorphism, monomorphism, and sequences in conserved microsatellites in primate species. Journal of Molecular Evolution 41 (4): 492-497

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends in Ecology & Evolution 18 (10): 503-511

Bogerd HP et al. (2006) Cellular inhibitors of long interspersed element 1 and *Alu* retrotransposition. Proceedings of the National Academy of Sciences of the United States of America 103 (23): 8780-5

Bowcock AM et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368 (6470): 455-7

Boyer JC et al. (2008) Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 640 (1-2): 89-96

Boyer JC et al. (2002) Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Human Molecular Genetics 11 (6): 707-713

Brandström M and Ellegren H (2008) Genome-wide analysis of microsatellite polymorphism circumventing the ascertainment bias. Genome Research gr.075242.107

Brinkmann B et al. (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. American Journal of Human Genetics 62 (6): 1408-15

Brock GJ et al. (1999) *Cis*-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. Human Molecular Genetics 8 (6): 1061-7

Brohede J and Ellegren H (1999) Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. Proceedings of the Royal Society of London Series B-Biological Sciences 266 (1421): 825-833

Brohede J et al. (2004) Individual variation in microsatellite mutation rate in barn swallows. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 545 (1-2): 73-80

Bull LN et al. (1999) Compound microsatellite repeats: practical and theoretical features. Genome Research 9 (9): 830-838

Buschiazzo E and Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28 (10): 1040-1050

Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. Journal of Forensic Sciences 51 (2): 253-265

Butler JT (2005) Forensic DNA Typing, Second. Elsevier Academic Press, London

Calabrese P and Durrett R (2003) Dinucleotide repeats in the *Drosophila* and human genomes have complex, length-dependent mutation processes. Molecular Biology and Evolution 20 (5): 715-25

Callen DF et al. (1993) Incidence and origin of "null" alleles in the $(AC)_n$ microsatellite markers. American Journal of Human Genetics 52 (5): 922-7

Chakraborty R et al. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. PNAS 94 (3): 1041-1046

Chambers GK and MacAvoy ES (2000) Microsatellites: consensus and controversy. Comp Biochem Physiol B Biochem Mol Biol 126 (4): 455-76

Chevet E et al. (1995) Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. Nucleic Acids Research 23 (16): 3343-3344

Chiaromonte F et al. (2001) Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. Proceedings of the National Academy of Sciences of the United States of America 98 (25): 14503-8

Clark RM et al. (2006) Expansion of GAA trinucleotide repeats in mammals. Genomics 87 (1): 57-67

Clark RM et al. (2004) Expansion of GAA triplet repeats in the human genome: unique origin of the FRDA mutation at the center of an *Alu*. Genomics 83 (3): 373-83

Clisson I et al. (2000) Conservation and evolution of microsatellite loci in primate taxa. American Journal of Primatology 50 (3): 205-214

Coghlan A and Durbin R (2007) Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure. Bioinformatics 23 (12): 1468-1475

Colson I and Goldstein DB (1999) Evidence for complex mutations at microsatellite loci in *Drosophila*. Genetics 152 (2): 617-627

Cooper G et al. (1999) Markov Chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. PNAS 96 (21): 11916-11921

Cooper G et al. (1998) Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. Human Molecular Genetics 7 (9): 1425-1429

Cooper GM et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Research 15 (7): 901-13

Corneo G et al. (1967) A satellite DNA isolated from human tissues. Journal of Molecular Biology 23 (3): 619-22

Costantini M et al. (2006) An isochore map of human chromosomes. Genome Research 16 (4): 536-541

Coulondre C et al. (1978) Molecular basis of base substitution hotspots in *Escherichia coli.* Nature 274 (5673): 775-80

Crawford AM et al. (1998) Microsatellite evolution: Testing the ascertainment bias hypothesis. Journal of Molecular Evolution 46 (2): 256-260

Cruz F et al. (2005) Distribution and abundance of microsatellites in the genome of bivalves. Gene 346 241-7

de Jong WW et al. (2003) Indels in protein-coding sequences of Euarchontoglires constrain the rooting of the euteriean tree. Molecular Phylogenetics and Evolution 28 328-340

Dermitzakis ET et al. (2005) Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nature Reviews Genetics 6 (2): 151-7

Dettman JR and Taylor JW (2004) Mutation and evolution of microsatellite loci in Neurospora. Genetics 168 (3): 1231-48

Di Rienzo A et al. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148 (3): 1269-1284

Di Rienzo A et al. (1994) Mutational processes of simple-sequence repeat loci in human populations. Proceedings of the National Academy of Sciences of the United States of America 91 (8): 3166-70

Dieringer D and Schlötterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Research 13 (10): 2242-2251

Domingo-Roura X (2002) Genetic distinction of marten species by fixation of a microsatellite region. Journal of Mammalogy 83 (3): 907-912

Domingo-Roura X et al. (2005) Phylogenetic inference and comparative evolution of a complex microsatellite and its flanking regions in carnivores. Genetical Research 85 (3): 223-233

Don RH et al. (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Research 19 (14): 4008

Duffy AJ et al. (1996) Microsatellites at a common site in the second ORF of L1 elements in mammalian genomes. Mammalian Genome 7 (5): 386-7

Dupuy BM et al. (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Human Mutation 23 (2): 117-24

Eckert KA and Yan G (2000) Mutational analyses of dinucleotide and tetranucleotide microsatellites in *Escherichia coli*: influence of sequence on expansion mutagenesis. Nucl. Acids Res. 28 (14): 2831-2838

Eddy SR (2005) A model of the statistical power of comparative genome sequence analysis. PLoS Biology 3 (1): e10

Ellegren H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. Nature Genetics 24 (4): 400-402

Ellegren H (2004) Microsatellites: Simple sequences with complex evolution. Nature Reviews Genetics 5 (6): 435-445

Ellegren H (2007) Characteristics, causes and evolutionary consequences of male-biased mutation. Proc Biol Sci 274 (1606): 1-10

Ellegren H et al. (2003) Mutation rate variation in the mammalian genome. Current Opinion in Genetics & Development 13 (6): 562-568

Estoup A et al. (1995) Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics 140 (2): 679-95

Estoup A et al. (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Molecular Ecology 11 (9): 1591-1604

Estoup A et al. (1993) Characterization of $(GT)_n$ and $(CT)_n$ microsatellites in two insect species: *Apis mellifera* and *Bombus terrestris*. Nucleic Acids Research 21 (6): 1427-31

Estoup A et al. (1995) Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). Molecular Biology and Evolution 12 (6): 1074-84

Ezenwa VO et al. (1998) Ancient conservation of trinucleotide microsatellite loci in polistine wasps. Molecular Phylogenetics and Evolution 10 (2): 168-177

Farber CR and Medrano JF (2004) Identification of putative homology between horse microsatellite flanking sequences and cross-species ESTs, mRNAs and genomic sequences. Animal Genetics 35 (1): 28-33

Ferguson-Smith MA and Trifonov V (2007) Mammalian karyotype evolution. Nature Reviews Genetics 8 (12): 950-962

Fink S et al. (2007) High variability and non-neutral evolution of the mammalian *avpr1* a gene BMC Evolutionary Biology 7 (176):

Fitzsimmons NN (1998) Single paternity of clutches and sperm storage in the promiscuous green turtle (Chelonia mydas). Molecular Ecology 7 (5): 575-584

FitzSimmons NN et al. (1995) Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Molecular Biology and Evolution 12 (3): 432-40

Fondon JW, III and Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. PNAS 101 (52): 18058-18063

Fullerton SM et al. (2001) Local rates of recombination are positively correlated with GC content in the human genome. Molecular Biology and Evolution 18 (6): 1139-1142

Gadberry MD et al. (2005) Primaclade - a flexible tool to find conserved PCR primers across multiple species. Bioinformatics 21 (7): 1263-4

Gardner MG et al. (2000) Microsatellite mutations in litters of the Australian lizard *Egernia stokesii*. Journal of Evolutionary Biology 13 (3): 551-560

Garza JC and Desmarais E (2000) Derivation of a simple microsatellite locus from a compound ancestor in the genus *Mus*. Mammalian Genome 11 (12): 1117-22

Garza JC et al. (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Molecular Biology and Evolution 12 (4): 594-603

Gemmell NJ et al. (1997) Interspecific microsatellite markers for the study of pinniped populations. Molecular Ecology 6 (7): 661-666

Gerstein MB et al. (2007) What is a gene, post-ENCODE? History and updated definition. Genome Research 17 (6): 669-81

Giardine B et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Research 15 (10): 1451-1455

Gibbs RA et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428 (6982): 493-521

Gill P et al. (1985) Forensic application of DNA 'fingerprints'. Nature 318 (6046): 577-9

Goldstein D and Clark A (1995) Microsatellite variation in North American populations of *Drosophila melanogaster*. Nucl. Acids Res. 23 (19): 3882-3886

Goldstein DB et al. (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139 (1): 463-71

Gonzalez-Martinez SC et al. (2004) Cross-amplification and sequence variation of microsatellite loci in Eurasian hard pines. Theoretical and Applied Genetics 109 (1): 103-111

Gordon AJ (1997) Microsatellite birth register. Journal of Molecular Evolution 45 (3): 337-8

Gow JL et al. (2005) A high incidence of clustered microsatellite mutations revealed by parent-offspring analysis in the African freshwater snail, *Bulinus forskalii* (Gastropoda, Pulmonata). Genetica 124 (1): 77-83

Grimaldi MC and Crouau-Roy B (1997) Microsatellite allelic homoplasy due to variable flanking sequences. Journal of Molecular Evolution 44 (3): 336-40

Grover A et al. (2007) Biased distribution of microsatellite motifs in the rice genome. Mol Genet Genomics

Gu L and Li GM (2006) Analysis of DNA mismatch repair in cellular response to DNA damage. Methods in Enzymology 408 303-17

Guillemaud T et al. (2000) Interspecific utility of microsatellites in fish: a case study of (CT)(n) and (GT)(n) markers in the shanny *Lipophrys pholis* (Pisces: Blenniidae) and their use in other Blennioidei. Mar Biotechnol (NY) 2 (3): 248-253

Gusmão L et al. (2005) Mutation rates at Y chromosome specific microsatellites. Human Mutation 26 (6): 520-528

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41 95-98

Hammock EAD and Young LJ (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308 (5728): 1630-1634

Hancock JM (1999) Microsatellites and other simple sequences: genomic context and mutational mechanisms. *In* D. GoldsteinC. Schlötterer, Oxford University Press, New York

Hancock JM and Simon M (2005) Simple sequence repeats in proteins and their significance for network evolution. Gene 345 (1): 113-8

Harr B and Schlötterer C (2000) Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics 155 (3): 1213-20

Harr B et al. (2002) Mismatch repair-driven mutational bias in *D. melanogaster*. Molecular Cell 10 (1): 199-205

Harr B et al. (2000) Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from *Drosophila*. Molecular Biology and Evolution 17 (7): 1001-9

Hawk JD et al. (2005) Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. Proceedings of the National Academy of Sciences of the United States of America 102 (24): 8639-43

Hecker KH and Roux KH (1996) High and low annealing temperatures increase both specificity and yield in touchdown and stepdown PCR. BioTechniques 20 (3): 478-85

Housley D et al. (2006) Design factors that influence PCR amplification success of cross-species primers among 1147 mammalian primer pairs. BMC Genomics 7 (1): 253

Huang QY et al. (2002) Mutation patterns at dinucleotide microsatellite loci in humans. American Journal of Human Genetics 70 (3): 625-34

Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17 (8): 754-5

Huson DH and Steel M (2004) Phylogenetic trees based on gene content. Bioinformatics 20 (13): 2044-2049

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431 (7011): 931-45

Jakupciak JP and Wells RD (2000) Gene conversion (recombination) mediates expansions of CTG·CAG repeats. Journal of Biological Chemistry 275 (51): 40003-13

Janke A et al. (1996) The mitochondrial genome of a monotreme--the platypus (*Ornithorhynchus anatinus*). Journal of Molecular Evolution 42 (2): 153-9

Janke A et al. (2002) Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: increased support for the Marsupionta hypothesis. Journal of Molecular Evolution 54 (1): 71-80

Jarne P and Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends in Ecology & Evolution 11 (10): 424-429

Jeffreys AJ et al. (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314 (6006): 67-73

Jin L et al. (1996) Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proceedings of the National Academy of Sciences of the United States of America 93 (26): 15285-8

Jin L et al. (1994) The exact numbers of possible microsatellite motifs. American Journal of Human Genetics 55 (3): 582-3

Jobb G (2007). TREEFINDER version of November 2007. Munich, Germany, Distributed by the author at www.treefinder.de.

Karhu A et al. (2000) Rapid expansion of microsatellite sequences in pines. Molecular Biology and Evolution 17 (2): 259-65

Karolchik D et al. (2003) The UCSC genome browser database. Nucleic Acids Research 31 (1): 51-54

Kashi Y and King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics 22 (5): 253-9

Katti MV et al. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution 18 (7): 1161-7

Kayser M et al. (2004) A comprehensive survey of human Y-chromosomal microsatellites. American Journal of Human Genetics 74 (6): 1183-1197

Kayser M et al. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. American Journal of Human Genetics 66 (5): 1580-8

Kayser M et al. (2006) Microsatellite length differences between humans and chimpanzees at autosomal loci are not found at equivalent haploid Y chromosomal loci. Genetics 173 (4): 2179-2186

Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution. Science 303 (5664): 1626-32

Kehrer-Sawatzki H and Cooper DN (2007) Structural divergence between the human and chimpanzee genomes. Human Genetics 120 (6): 759-78

Kehrer-Sawatzki H and Cooper DN (2007) Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. Human Mutation 28 (2): 99-130

Kelkar YD et al. (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Research 18 30-38

Kent WJ et al. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences of the United States of America 100 (20): 11484-11489

Kim K-S et al. (2004) Cross-species amplification of Bovidae microsatellites and low diversity of the endangered Korean goral. Journal of Heredity 95 (6): 521-525

Kim T-S et al. (2008) Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics 9 (1): 31

Kimura M and Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49 725-38

King DC et al. (2007) Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. Genome Research 17 (6): 775-86

King DG and Kashi Y (2007) Mutability and evolvability: Indirect selection for mutability. Heredity 99 (2): 123-124

Klintschar M et al. (2004) Haplotype studies support slippage as the mechanism of germline mutations in short tandem repeats. Electrophoresis 25 (20): 3344-3348

Kofler R et al. (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23 (13): 1683-1685

Kondo Y et al. (1993) DNA segments mapped by reciprocal use of microsatellite primers between mouse and rat. Mammalian Genome 4 (10): 571-6

Kovalchuk O et al. (2003) Extremely complex pattern of microsatellite mutation in the germline of wheat exposed to the post-Chernobyl radioactive contamination. Mutation Research 525 (1-2): 93-101

Kovtun IV and McMurray CT (2008) Features of trinucleotide repeat instability in vivo. Cell Research 18 (1): 198-213

Kriegs JO et al. (2006) Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biology 4 (4):

Kruglyak S et al. (2000) Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Molecular Biology and Evolution 17 (8): 1210-1219

Kruglyak S et al. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proceedings of the National Academy of Sciences of the United States of America 95 (18): 10774-10778

Kumar S and Filipski A (2007) Multiple sequence alignment: In pursuit of homologous DNA positions. Genome Research 17 (2): 127-135

La Rota M et al. (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6 (1): 23

Lai Y and Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. Molecular Biology and Evolution 20 (12): 2123-31

Laidlaw J et al. (2007) Elevated basal slippage mutation rates among the Canidae. Journal of Heredity 98 (5): 452-460

Lander ES et al. (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822): 860-921

Lander ES and Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2 (3): 231-9

Lawson M and Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biology 7 (2): R14

Leclercq S et al. (2007) Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinformatics 8 (1): 125

Lee JS et al. (1999) Relative stabilities of dinucleotide and tetranucleotide repeats in cultured mammalian cells. Human Molecular Genetics 8 (13): 2567-72

Lei X et al. (2004) Measurement of DNA mismatch repair activity in live cells. Nucleic Acids Research 32 (12): e100

Leopoldino AM and Pena SD (2002) The mutational spectrum of human autosomal tetranucleotide microsatellites. Human Mutation 21 (1): 71-9

Levinson G and Gutman GA (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Research 15 (13): 5323-38

Li G-M (2008) Mechanisms and functions of DNA mismatch repair. Cell Research 18 (1): 85-98

Li YC et al. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11 (12): 2453-65

Li YC et al. (2004) Microsatellites within genes: structure, function, and evolution. Molecular Biology and Evolution 21 (6): 991-1007

Lia V et al. (2007) Complex mutational patterns and size homoplasy at maize microsatellite loci. Theoretical and Applied Genetics 115 (7): 981-991

Liewlaksaneeyanawin C et al. (2004) Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theoretical and Applied Genetics 109 (2): 361-9

Lindblad-Toh K et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438 (7069): 803-19

Lipatov M et al. (2006) A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. Journal of Molecular Evolution 62 (2): 168-75

Litt M and Luty JA (1989) A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. American Journal of Human Genetics 44 (3): 397-401

Loots G and Ovcharenko I (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. Bioinformatics 23 (1): 122-4

Lopez-Giraldez F et al. (2007) High Incidence of nonslippage mechanisms generating variability and complexity in eurasian badger microsatellites. Journal of Heredity 98 (6): 620-628

Lowe CB et al. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. PNAS 104 (19): 8005-8010

Lunter G et al. (2006) Genome-wide identification of human functional DNA using a neutral indel model. PLoS Computational Biology 2 (1): e5

Luo S-J et al. (2007) Development of Y chromosome intraspecific polymorphic markers in the Felidae. Journal of Heredity esm063

MacDonald AJ et al. (2006) Y chromosome microsatellite markers identified from the tammar wallaby (*Macropus eugenii*) and their amplification in three other macropod species. Molecular Ecology Notes 6 1202-1204

Makova KD et al. (2000) Evolution of microsatellite alleles in four species of mice (genus *Apodemus*). Journal of Molecular Evolution 51 (2): 166-172

Marcotte EM et al. (1999) A census of protein repeats. Journal of Molecular Biology 293 (1): 151-60

Margulies EH and Birney E (2008) Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. Nature Reviews Genetics 9 303-313

Margulies EH et al. (2006) Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends in Genetics 22 (4): 187-93

Margulies EH et al. (2007) Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Research 17 (6): 760-774

Martin AP et al. (2002) Conservation of a dinucleotide simple sequence repeat locus in sharks. Molecular Phylogenetics and Evolution 23 (2): 205-213

McConnell R et al. (2007) An unusually low microsatellite mutation rate in *Dictyostelium discoideum*, an organism with unusually abundant microsatellites. Genetics genetics.107.076067

Messier W et al. (1996) The birth of microsatellites. Nature 381 (6582): 483

Metzgar D et al. (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Research 10 (1): 72-80

Metzgar D et al. (2002) Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. Genome Research 12 (3): 408-413

Meyer E et al. (1995) Microsatellite polymorphisms reveal phylogenetic relationships in primates. Journal of Molecular Evolution 41 (1): 10-4

Mikkelsen TS et al. (2007) Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature 447 (7141): 167-177

Mikul et al. (2007) Can microsatellite markers resolve phylogenetic relationships between closely related crested newt species (*Triturus cristatus* superspecies)? Amphibia-Reptilia 28 467-474

Miller W et al. (2004) Comparative genomics. Annual Review of Genomics and Human Genetics 5 (1): 15-56

Miller W et al. (2007) 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Research 17 1797-1808

Mills RE et al. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research 16 (9): 1182-1190

Mirkin SM (2007) Expandable DNA repeats and human disease. Nature 447 (7147): 932-940

Moore SS et al. (1998) NCAM: a polymorphic microsatellite locus conserved across eutherian mammal species. Animal Genetics 29 (1): 33-36

Moore SS et al. (1991) The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. Genomics 10 (3): 654-660

Morgante M et al. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genetics 30 (2): 194-200

Mrowka R et al. (2007) Dissecting the action of an evolutionary conserved non-coding region on renin promoter activity. Nucl. Acids Res. 35 (15): 5120-5129

Murphy WJ and O'Brien SJ (2007) Designing and optimizing comparative anchor primers for comparative gene mapping and phylogenetic inference. Nature Protocols 2 (11): 3022-3030

Nadir E et al. (1996) Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proceedings of the National Academy of Sciences of the United States of America 93 (13): 6470-5

Neff BD and Gross MR (2001) Microsatellite evolution in vertebrates: Inference from AC dinucleotide repeats. Evolution 55 (9): 1717-1733

Nevo E et al. (2005) Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel. Biological Journal of the Linnean Society 84 (2): 205-224

Nikitina TV et al. (2005) Germline mutations of tetranucleotide DNA repeats in families with normal children and reproductive pathology. Russian Journal of Genetics 41 (7): 770-778

Nikolaev S et al. (2007) Early history of mammals Is elucidated with the ENCODE multiple species sequencing data. PLoS Genet 3 (1): e2

Nishizawa M and Nishizawa K (2002) A DNA sequence evolution analysis generalized by simulation and the Markov chain Monte Carlo method implicates strand slippage in a majority of insertions and deletions. Journal of Molecular Evolution 55 (6): 706-717

Noor MA and Feder JL (2006) Speciation genetics: evolving approaches. Nature Reviews Genetics 7 (11): 851-61

Noor MA et al. (2001) Evolutionary history of microsatellites in the obscura group of Drosophila. Molecular Biology and Evolution 18 (4): 551-6

O'Dushlaine C et al. (2005) Tandem repeat copy-number variation in protein-coding regions of human genes. Genome Biology 6 (8): R69

Ota T and Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetical Research 22 (2): 201-4

Paetkau D and Strobeck C (1995) The molecular basis and evolutionary history of a microsatellite null allele in bears. Molecular Ecology 4 (4): 519-20

Pagel M and Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systems Biology 53 (4): 571-81

Pagel M et al. (2004) Bayesian estimation of ancestral character states on phylogenies. Systematic Biology 53 (5): 673-684

Palo JU et al. (2001) Microsatellite variation in ringed seals (*Phoca hispida*): genetic structure and history of the Baltic Sea population. Heredity 86 (Pt 5): 609-17

Pardi F et al. (2005) On the structural differences between markers and genomic AC microsatellites. Journal of Molecular Evolution 60 (5): 688-693

Parida SK et al. (2006) Unigene derived microsatellite markers for the cereal genomes. Theoretical and Applied Genetics 112 (5): 808-17

Park MH et al. (2008) Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean. Journal of Human Genetics 53 (3): 254-266

Pashley CH et al. (2006) EST databases as a source for molecular markers: Lessons from *Helianthus*. Journal of Heredity 97 (4): 381-388

Paszek AA et al. (1998) Evaluating evolutionary divergence with microsatellites. Journal of Molecular Evolution 46 (1): 121-6

Pavlov YI et al. (2003) Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. Current Biology 13 (9): 744-8

Pearson CE et al. (2005) Repeat instability: mechanisms of dynamic mutations. Nature Reviews Genetics 6 (10): 729-42

Pennacchio LA et al. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science 294 (5540): 169-73

Perez F et al. (2005) Development of EST-SSR markers by data mining in three species of shrimp: *Litopenaeus vannamei*, *Litopenaeus stylirostris*, and *Trachypenaeus birdy*. Marine Biotechnology (NY) 7 (5): 554-69

Perez M et al. (2005) Distribution properties of polymononucleotide repeats in molluscan genomes. Journal of Heredity 96 (1): 40-51

Petes TD et al. (1997) Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. Genetics 146 (2): 491-8

Pheasant M and Mattick JS (2007) Raising the estimate of functional human sequences. Genome Research 17 (9): 1245-1253

Prabhakar S et al. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Research 16 (7): 855-63

Prakash A and Tompa M (2007) Measuring the accuracy of genome-size multiple alignments. Genome Biology 8 (6): R124

Primmer CR and Ellegren H (1998) Patterns of molecular evolution in avian microsatellites. Molecular Biology and Evolution 15 (8): 997-1008

Primmer CR et al. (1996) Directional evolution in germline microsatellite mutations. Nature Genetics 13 (4): 391-393

Primmer CR et al. (1996) A wide-range survey of cross-species microsatellite amplification in birds. Molecular Ecology 5 (3): 365-378

Primmer CR et al. (1997) Low frequency of microsatellites in the avian genome. Genome Research 7 (5): 471-82

Primmer CR et al. (1998) Unraveling the processes of microsatellite evolution through analysis of germ line mutations in barn swallows *Hirundo rustica*. Molecular Biology and Evolution 15 (8): 1047-1054

Primmer RC et al. (2005) Factors affecting avian cross-species microsatellite amplification. Journal of Avian Biology 36 (4): 348-360

Pumpernik D et al. (2008) Replication slippage versus point mutation rates in short tandem repeats of the human genome. Molecular Genetics and Genomics 279 (1): 53-61

Pupko T and Graur D (1999) Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units. Journal of Molecular Evolution 48 (3): 313-6

Rambaut A (2006-2008). FigTree v1.1.

Raveendran M et al. (2006) Designing new microsatellite markers for linkage and population genetic analyses in rhesus macaques and other nonhuman primates. Genomics 88 (6): 706-710

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316 (222): 222-234

Rice P et al. (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics 16 (6): 276-277

Richard GF and Paques F (2000) Mini- and microsatellite expansions: the recombination connection. EMBO Reports 1 (2): 122-6

Richard M and Thorpe RS (2001) Can microsatellites be used to infer phylogenies? Evidence from population affinities of the western Canary Island lizard (*Gallotia galloti*). Molecular Phylogenetics and Evolution 20 (3): 351-360

Rico C et al. (1996) 470 million years of conservation of microsatellite loci among fish species. Proceedings of the Royal Society of London B Biological Sciences 263 (1370): 549-557

Riley DE et al. (2007) Simple repeat evolution includes dramatic primary sequence changes that conserve folding potential. Biochemical and Biophysical Research Communications 355 (3): 619-625

Riley DE and Krieger JN (2004) Simple repeat replacements support similar functions of distinct repeats in inter-species mRNA homologs. Gene 328 17-24

Riley DE and Krieger JN (2005) Short tandem repeat (STR) replacements in UTRs and introns suggest an important role for certain STRs in gene expression and disease. Gene 344 203-11

Robert VJ et al. (2005) Chromatin and RNAi factors protect the *C. elegans* germline against repetitive sequences. Genes & Development 19 (7): 782-7

Robertson BC and Gemmell NJ (2004) Defining eradication units to control invasive pests. Journal of Applied Ecology 41 (6): 1042-1048

Roest Crollius H et al. (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. Nature Genetics 25 (2): 235-8

Rolfsmeier ML et al. (2000) Mismatch repair blocks expansions of interrupted trinucleotide repeats in yeast. Molecular Cell 6 (6): 1501-7

Rolfsmeier ML and Lahue RS (2000) Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae*. Molecular and Cellular Biology 20 (1): 173-80

Ronquist F and Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19 (12): 1572-4

Rose O and Falush D (1998) A threshold size for microsatellite expansion. Molecular Biology and Evolution 15 (5): 613-5

Ross CL et al. (2003) Rapid divergence of microsatellite abundance among species of *Drosophila*. Molecular Biology and Evolution 20 (7): 1143-1157

Rossiter SJ et al. (2007) Rangewide phylogeography in the greater horseshoe bat inferred from microsatellites: implications for population history, taxonomy and conservation. Molecular Ecology 16 (22): 4699-4714

Rout P et al. (2008) Microsatellite-based phylogeny of Indian domestic goats. BMC Genetics 9 (1): 11

Rozen S and Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology 132 365-86

Rubinsztein DC et al. (1995) Microsatellite evolution - evidence for directionality and variation in rate between species. Nature Genetics 10 (3): 337-343

Ruiz-Garcia M (2005). The use of several microsatellite loci applied to 13 neotropical primates revealed a strong recent bottleneck event in the woolly monkey (*Lagothrix lagotricha*) in Colombia. Primate Report (71). M. H. Schwibbe. Goettingen, German Primate Centre**:** 27-55.

Sainudiin R et al. (2004) Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. Genetics 168 (1): 383-395

Santibáñez-Koref MF et al. (2001) A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. Molecular Biology and Evolution 18 (11): 2119-23

Schlötterer C (2004) The evolution of molecular markers - just a matter of fashion? Nature Reviews Genetics 5 (1): 63-9

Schlötterer C et al. (1991) Conservation of polymorphic simple sequence loci in cetacean species. Nature 354 (6348): 63-5

Schlötterer C et al. (1998) High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. Molecular Biology and Evolution 15 (10): 1269-1274

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nature Biotechnology 18 (2): 233-4

Schug MD et al. (1998) The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. Molecular Biology and Evolution 15 (12): 1751-60

Schug MD et al. (1997) Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nature Genetics 15 (1): 99-102

Shao Z et al. (2005) Complex mutation at a microsatellite locus in sturgeons: *Acipenser sinensis*, *A. schrenckii*, *A. gueldenstaedtii* and *A. baerii*. Journal of Applied Ichthyology 21 (1): 2-6

Sharma PC et al. (2007) Mining microsatellites in eukaryotic genomes. Trends in Biotechnology

Shepherd LD and Lambert DM (2005) Mutational bias in penguin microsatellite DNA. Journal of Heredity 96 (5): 566-571

Shinde D et al. (2003) *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. Nucleic Acids Research 31 (3): 974-80

Shriver MD et al. (1995) A novel measure of genetic-distance for highly polymorphic tandem repeat loci. Molecular Biology and Evolution 12 (5): 914-920

Sia EA et al. (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Molecular and Cellular Biology 17 (5): 2851-8

Sibly RM et al. (2003) The structure of interrupted human AC microsatellites. Molecular Biology and Evolution 20 (3): 453-459

Sibly RM et al. (2001) A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. Molecular Biology and Evolution 18 (3): 413-417

Siddappa NB et al. (2007) Regeneration of commercial nucleic acid extraction columns without the risk of carryover contamination. BioTechniques 42 (2): 186-192

Siepel A et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15 (8): 1034-50

Simons C et al. (2007) Maintenance of transposon-free regions throughout vertebrate evolution. BMC Genomics 8 (1): 470

Simons C et al. (2006) Transposon-free regions in mammalian genomes. Genome Research 16 (2): 164-72

Slate J et al. (1998) Bovine microsatellite loci are highly conserved in red deer (*Cervus elaphus*), sika deer (*Cervus nippon*) and Soay sheep (*Ovis aries*). Animal Genetics 29 (4): 307-315

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139 (1): 457-62

Smit A et al. (1996-2007). Repeat-Masker Open-3.0.

Smith NGC et al. (2004) Evidence for turnover of functional noncoding DNA in mammalian genome evolution. Genomics 84 (5): 806-813

Stallings RL (1995) Conservation and evolution of $(CT)_n/(GA)_n$ microsatellite sequences at orthologous positions in diverse mammalian genomes. Genomics 25 (1): 107-113

Stefanini FM and Feldman MW (2000) Bayesian estimation of range for microsatellite loci. Genetical Research 75 (2): 167-77

Steiper ME and Young NM (2006) Primate molecular divergence dates. Molecular Phylogenetics and Evolution 41 (2): 384-394

Stephan W and Cho S (1994) Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics 136 (1): 333-41

Stephan W and Kim Y (1998) Persistence of microsatellite arrays in finite populations. Molecular Biology and Evolution 15 (10): 1332-1336

Sturzeneker R et al. (2000) Microsatellite instability in tumors as a model to study the process of microsatellite mutations. Human Molecular Genetics 9 (3): 347-52

Subirana JA and Messeguer X (2008) Structural families of genomic microsatellites. Gene 408 (1-2): 124-132

Subramanian S et al. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biology 4 (2): R13

Sun HS and Kirkpatrick BW (1996) Exploiting dinucleotide microsatellites conserved among mammalian species. Mammalian Genome 7 (2): 128-132

Sunnucks P (2000) Efficient genetic markers for population biology. Trends in Ecology & Evolution 15 (5): 199-203

Swofford DL (2002). PAUP*: Phylogenetic Analyses Using Parsimony* (and Other Methods). Sinauer. Sunderland, MA.

Tachida H and Iizuka M (1992) Persistence of repeated sequences that evolve by replication slippage. Genetics 131 (2): 471-8

Tamiya G et al. (2005) Whole genome association study of rheumatoid arthritis using 27,039 microsatellites. Human Molecular Genetics 14 (16): 2305-2321

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucl. Acids Res. 17 (16): 6463-6471

Taylor JS et al. (1999) The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Molecular Biology and Evolution 16 (4): 567-72

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447 (7146): 799-816

Thompson JD et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22 (22): 4673-80

Thurman RE et al. (2007) Identification of higher-order functional domains in the human ENCODE regions. Genome Research 17 (6): 917-927

Tóth G et al. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research 10 (7): 967-81

Udupa SM et al. (2004) Tightly linked di- and tri-nucleotide microsatellites do not evolve in complete independence: evidence from linked $(TA)_n$ and $(TAA)_n$ microsatellites of chickpea (*Cicer arietinum* L.). Theoretical and Applied Genetics 108 (3): 550-557

Ustinova J et al. (2006) Long repeats in a huge genome: microsatellite loci in the grasshopper *Chorthippus biguttulus*. Journal of Molecular Evolution 62 (2): 158-167

van Belkum A (2007) Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA). FEMS Immunology; Medical Microbiology 49 22-27

van Oppen MJH et al. (2000) Extensive homoplasy, nonstepwise mutations, and shared ancestral polymorphism at a complex microsatellite locus in Lake Malawi cichlids. Molecular Biology and Evolution 17 (4): 489-498

van Rheede T et al. (2006) The platypus is in its place: Nuclear genes and indels confirm the sister group relation of monotremes and therians. Molecular Biology and Evolution 23 (3): 587-597

Vardhanabhuti S et al. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Research 35 (10): 3203-13

Vargas-Jentzsch I et al. (2008) Evolution of microsatellite DNA. *In* L. John Wiley & Sons, Chichester

Varshney RK et al. (2005) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Science 168 (1): 195-202

Venkatesh B et al. (2006) Ancient noncoding elements conserved in the human genome. Science 314 (5807): 1892

Vigouroux Y et al. (2002) Rate and pattern of mutation at microsatellite loci in maize. Molecular Biology and Evolution 19 (8): 1251-60

Vigouroux Y et al. (2003) Directional evolution for microsatellite size in maize. Molecular Biology and Evolution 20 (9): 1480-1483

Vogt P (1990) Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". Human Genetics 84 (4): 301-36

Vowles EJ and Amos W (2006) Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. Molecular Biology and Evolution 23 (3): 598-607

Walsh PW et al. (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. BioTechniques 10 (4): 506-513

Warren WC et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. Nature 453 (7192): 175-183

Waterston RH et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420 (6915): 520-62

Weber J and Wong C (1993) Mutation of human short tandem repeats. Human Molecular Genetics 2 (8): 1123-1128

Weber JL (1990) Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7 (4): 524-30

Webster MT et al. (2002) Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proceedings of the National Academy of Sciences of the United States of America 99 (13): 8748-8753

Weetman D et al. (2002) Reconstruction of microsatellite mutation history reveals a strong and consistent deletion bias in invasive clonal snails, *Potamopyrgus antipodarum*. Genetics 162 (2): 813-22

Weissenbach J et al. (1992) A second-generation linkage map of the human genome. Nature 359 (6398): 794-801

Whitham TG et al. (2006) A framework for community and ecosystem genetics: from genes to ecosystems. Nature Reviews Genetics 7 (7): 510-23

Whittaker JC et al. (2003) Likelihood-based estimation of microsatellite mutation rates. Genetics 164 (2): 781-787

Wiehe T et al. (2007) Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. Genetics 175 (1): 207-218

Wierdl M et al. (1997) Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146 (3): 769-779

Wilder J and Hollocher H (2001) Mobile elements and the genesis of microsatellites in dipterans. Molecular Biology and Evolution 18 (3): 384-392

Wildman DE et al. (2007) Genomics, biogeography, and the diversification of placental mammals. Proceedings of the National Academy of Sciences of the United States of America 104 (36): 14395-14400

Woerner SM et al. (2006) Microsatellite instability in the development of DNA mismatch repair deficient tumors. Cancer Biomark 2 (1-2): 69-86

Wong KM et al. (2008) Alignment uncertainty and genomic analysis. Science 319 (5862): 473-476

Woolfe A et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biology 3 (1): e7

Wren JD et al. (2000) Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. American Journal of Human Genetics 67 (2): 345-56

Xie X et al. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. PNAS 104 (17): 7145-7150

Xu H et al. (2005) Mutation rate variation at human dinucleotide microsatellites. Genetics 170 (1): 305-12

Xu X et al. (2000) The direction of microsatellite mutations is dependent upon allele length. Nature Genetics 24 (4): 396-9

Yamada NA et al. (2002) Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 499 (2): 213-225

Yauk CL and Polyzos A (2005) Tandem repeat DNA: applications in mutation analysis. Environmental Mutagen Research 27 93-98

Yoder JA et al. (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends in Genetics 13 (8): 335-40

Zhang L et al. (2004) Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. Bioinformatics 20 (7): 1081-6

Zhu Y et al. (2000) A phylogenetic perspective on sequence evolution in microsatellite loci. Journal of Molecular Evolution 50 (4): 324-338

Zhu Y et al. (2000) Insertions, substitutions, and the origin of microsatellites. Genetical Research 76 (3): 227-36

Ziegele JS et al. (1992) Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics 14 1026-1031